THE SEQUENCE SPECIFIC DNA BINDING ACTIVITY OF *POGO* TRANSPOSASE

Hongmei Wang

A thesis presented for the degree of PhD

University of Edinburgh

1997

I dedicate this thesis to
my husband Chongbo, my daughter Mengyuan, and my parents.

DECLARATION

This thesis and all the work herein was composed by myself, unless otherwise stated.

Hongmei Wang
April, 1997

ACKNOWLEDGEMENTS

First of all, I would like to appreciate my supervisor Professor David Finnegan for his patient, multiple supervisions during my studies - from science to language. I will definitely benefit from the knowledge of molecular biology I've learned from his lab, however, more importanly, I've learned how to conduct research scientifically, and how to deal with the success and unsuccess in the research process. I would also like to appreciate my second supervisor, Professor Ken Murray for his advice.

Many thanks go to people who is working and used to worked in the Finnegan group, Angela, Apinunt, Eve, Ivan, Jill, Julie, Laura, Maki, Petra and Susan, particularly my labmates Petra and Ruth, for their help, encouragement and friendship. And people from 6th floor Darwin. I really enjoy the time to be with them.

I would like to thank Dr. David Dryden for his advice on my protein mutagenesis work.

Thanks should also go to my Chinese friends in ICMB, Xinsheng, Debiao, Suling and Wumin, for their friendship and chatting with me in our mother language. I would also like to thank my colleagues and friends in Nankai university, China, for their encouragement and friendship from my hometown.

Finally, I would like to thank Darwin Trust for supporting my PhD studies.

ABBREVIATIONS

A	adenosine
Amp	ampicillin
ATP	alkaline phosphatase
bp	base pair
BSA	bovine serum albumin
C	cytidine
°C	degree celsius
cDNA	complementary DNA
Ci	curie
cpm	counts per minute
d	deoxy
dd	dideoxy
DMSO	dimethhl sulphoxide
DNase	deoxyribonuclease
dNTP	deoxyribonucleoside triphosphate
DTT	ditiothreitol
EDTA	diaminoethanetetra-acetic acid
g	gram
G	guanosine
GST	glutathione S-transferase
hr	hour
HEPES	N-[2-Hydroxyethyl]piperazine-N'-[2-ethane-sulphonic acid]
IPTG	isopropyl-β-D-thiogalacopyranoside
kb	kilobase
kD	kiloDalton
l	litre
L-broth	Luria broth
M	molar
mg	milligram
min	minute
ml	millilitre
mol	mole
MOPS	3-[N-Morpholino]propanesulphonic acid
MW	molecular weight
ng	nanogram

O. D.	optical density
p	plasmid
p	pico
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PEG	polyethlene glycol
PIPES	piperazine-N, N'-bis [2-ethane-sulphonic acid]; 1,4-piperazine
POD	peroxidase
PVDF	polyvinylidene fluoride
RNase	ribonuclease
rpm	revolutions per minute
sec	second
SDS	sodium dodecyl sulphate
T	thymidine
TEMED	N,N,N'-trtramethylethylenediamine
T_m	melting temperature
Tris	tris (hydroxymethyl) aminomethane
U	uridine
µg	microgram
µl	microlitre
UV	ultraviolet
v	volt
v	volume
v/v	volume per volume
w/v	weight per volume
w.t.	wild type
X-gal	5-bromo-4-chloro-3-indol-β-galactopyranoside.

AMINO ACIDS

Amino acids	Three letter abbreviation	One letter symbol
alanine	Ala	A
arginine	Arg	R
asparagine	Asn	N
aspartic acid	Asp	D
cysteine	Cys	C
glutamine	Gln	Q
glutamic acid	Glu	E
glycine	Gly	G
histidine	His	H
isoleucine	Ile	I
leucine	Leu	L
lysine	Lys	K
methionine	Met	M
phenylalanine	Phe	F
proline	Pro	P
serine	Ser	S
threonine	Thr	T
tryptophan	Try	W
tyrosine	Tyr	Y
valine	Val	V

ABSTRACT

Transposable elements are DNA fragments which move from one position to another within and sometimes between genomes. They have been found in organisms ranging from bacteria to human, and have been shown making up 10-20 % of the total genome in eukaryotes. Transposase is the protein encoded by the transposon which is responsible for catalysing transposition reactions.

pogo is a transposable element in *Drosophila melanogaster* which transposes from DNA to DNA. It is 2.1 kb in size with 21 bp terminal inverted repeats and two open reading frames. cDNA analysis shows the two ORFs code for a single peptide which is supposed to be the putative transposase. However, no more information was available about this transposase before the start of this thesis.

In this thesis, the two ORFs believed to code for *pogo* transposase were joined in frame by PCR and cloned into a pGEX vector. The *pogo* transposase was fused to the C-termini of glutathione-S-transferase (GST) and expressed in *E. coli.*

The DNA binding activity of *pogo* transposase was tested by gel retardation assays in the presence of specific and non-specific competitors. *pogo* transposase was shown to be able to bind specifically to the end sequences of the element.

The transposase binding sites within the *pogo* element were identified by testing the transposase binding ability of deleted end sequences of the element. A 12 base pair consensus sequence was found and shown to be responsible for binding to the transposase. There are several copies of the 12 bp transposase binding site located near each end, in the 5' subterminal region, and in the middle of the element respectively. The binding sites at the ends might be involved in forming the synaptic DNA-protein complex and others in regulating transposition.

The DNA binding domain of *pogo* transposase was identified by expressing different regions of the transposase and determining their sequence specific DNA binding ability. The DNA binding domain has been shown to be located in the N-terminal 75 amino acid region. Site directed mutagenesis was used to study the role of the predicted helix-turn-helix (HTH) motif located within the DNA binding domain. Substitution of the positively charged basic amino acids within the recognition helix by alanine abolished the DNA binding activity of the protein; Substitutions introducing prolines into the first or second helices to disrupt their structures also greatly reduced the DNA binding activity of the protein. These data support the idea that the HTH motif in the DNA binding domain is responsible for the specific DNA binding activity of the transposase.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1.1 Transposable elements and transposition

Transposable elements are pieces of double stranded DNA that move from one position to another within and sometimes between genomes. They have been detected in organisms ranging from bacteria to humans (Berg and Howe, 1989), and can be divided into two main classes according to their structure and presumed mechanism of transposition (Finnegan, 1992) (Fig.1.1-1&2). Class I elements transpose by retrotransposition, in which the transposon is transcribed to an RNA intermediate prior to synthesis of a DNA copy. Class II elements transpose by transposition, in which DNA is the only intermediate. Prokaryotic organisms only have class II elements while eukaryotes have both class I and class II elements (Berg and Howe, 1989).

1.1.1 Class I elements and retrotransposition

In retrotransposition, the integrated DNA copy of the element is first transcribed into the RNA intermediate by host RNA polymerase, followed by reverse transcription of the RNA into a double-stranded DNA by the element-encoded reverse transcriptase. Integration of the resulting DNA copy into a suitable target DNA molecule is catalysed by the element-encoded integrase (IN) (Fig1.1-1). The best characterised class I elements are retroviruses and retrotransposons.

Retrotransposons represent a great variety of elements, from the ones that resemble a retrovirus in structure and mode of transposition (LTR retrotransposons) to simple sequences that seem to rely strictly on enzymes encoded elsewhere operating in *trans* [Short interspersed nuclear elements (SINEs)]. The elements that encode reverse transcriptase are divided into two major groups according to their structure and transposition mechanisms. The LTR elements are those containing long terminal repeats, and the non-LTR elements [LINEs (long interspersed nuclear elements)] are those without long terminal repeats.

LTR elements

LTR elements are normally 4.5-9.0 kb in size with long direct terminal repeats of 200-500 bp. They have two or three open reading frames encoding the proteins necessary for replication and transposition of the elements. These ORFs are named *gag*, *pol* and *env* after similar retroviral ORFs. The protein encoded by *gag* is cleaved into smaller matrix, capsid, and nucleocapsid proteins which are responsible for the protein-nucleic acid structure necessary for reverse transcription. The protein encoded by *pol* is cleaved to generate reverse transcriptase, aspartic protease, RNase H, and integrase. The final ORF, *env*, encodes components of the viral envelope that mediate entry of the virus into the host cell.

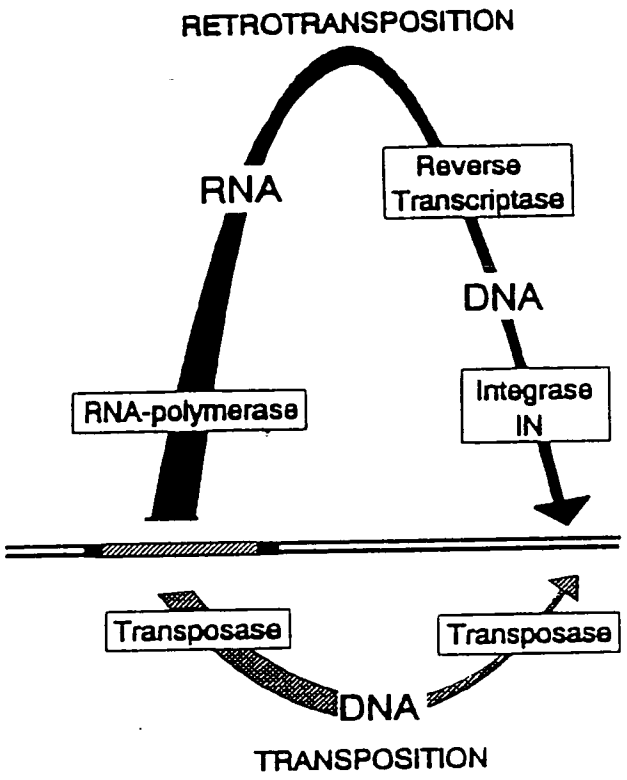


Fig. 1.1-1 Retrotransposition and transposition

After Polard and Chandler (1995)

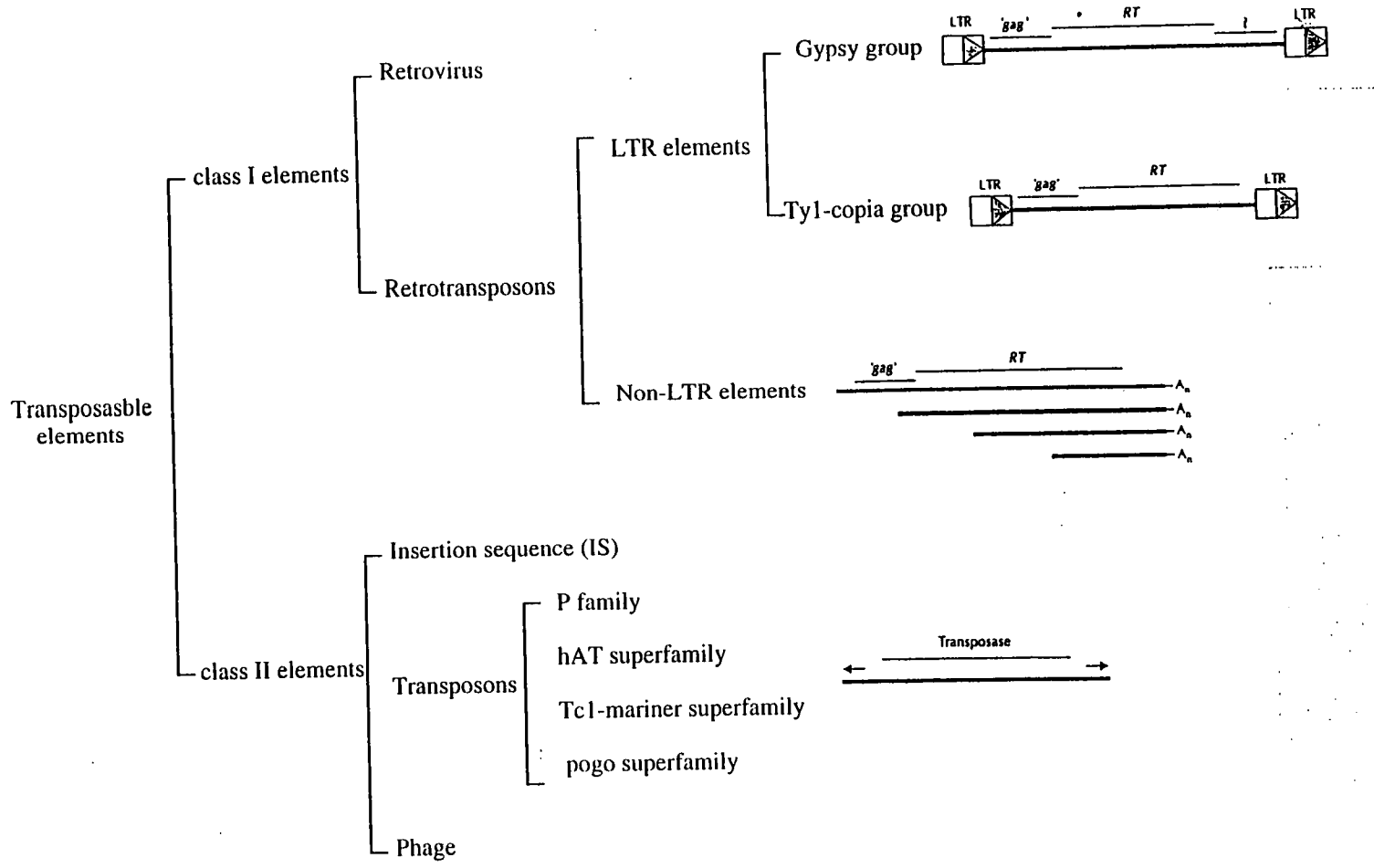


Fig.1.1-2 Classification of transposable elements

(after Finnegan, 1992 and Polard & Chandler, 1995)

LTR elements can be classified into two groups according to the numbers of their ORF. *Ty1-copia* group contains most LTR elements which only have two ORFs with the potential to code for polypeptides similar to the *gag* and *pol* encoded proteins, while *gypsy* group contains elements more like retroviruses in having a third ORF coding for a protein similar to a virus *env* protein. The *Ty1* elements of *Saccharomyces cerevisiae* and the *gypsy* elements of *Drosophila melanogaster* are examples of these two groups.

Ty 1 elements

Ty1 from *Saccharomyces cerevise* is a 5.9 kb element with 338 bp LTRs. It has two ORFs named TYA1 and TYB1 respectively. It has been shown that the 58 kd TYA1 protein is processed by *Ty1* protease within the virus like particle (VLP) to produce a 54 kd protein which is the structural protein of the VLP. The 190 kd TYB1 protein is processed in a similar way to generate a 23 kd protease, a 90 kd integrase and a 60 kd reverse transcriptase/RNaseH (Belcourt & Farabaugh, 1990; Adams et al, 1987; Garfinkel et al, 1991; Muller et al, 1987; Youngren et al. 1988). The reverse transcription of *Ty1* RNA is also supposed to be associated with the VLP (Garfinkel et al, 1985; Mellor et al, 1985). The *Ty1* cDNA is reintegrated into the host genome using the VLP as an intermediate (Eichinger & Boeke, 1988).

gypsy elements

The *gypsy* element from *Drosophila melanogaster* is 7.5 kb in size and has 486 bp long terminal repeats at both ends (Bayer et al., 1984, Freund & Meselson, 1984; Marlor et al., 1986). In addition to the two ORFs which resemble retroviral *gag* and *pol* genes, it has a third ORF which is of similar size and located in a position equivalent to the *env* genes of retroviruses, and contains a putative membrane-spanning domain near the C-terminus. The protein encoded by ORF3 is glycosylated and processed like all retroviral envelop proteins. The *env* protein is expressed at high levels in fly strains in which *gypsy* elements are active. *gypsy* has been suggested to be an infectious insect retrovirus since high level of *gypsy* insertion activity is observed when fly larvae are exposed to the viral particles (Song et al., 1994; Kim et al., 1994).

Non-LTR elements (LINE elements)

Non-LTR elements, also called LINEs (long interspersed nuclear elements) are repeated sequences that are widespread in many eukaryotic species. They include *I*, *F*, *Doc*, and *jockey* in *Drosophila melanogaster* (Fawcett et al., 1986; Di Nocera &

Casari, 1987; O'Hare et al., 1991; Priimagi et al., 1988), *ingi* in *Trypanosoma brucei* (Kimmel et al., 1987), *cin-4* in *Zea mays* (Schwarz-Sommer et al., 1987), and *L1* elements in mammals. They are devoid of terminal repeats but contain an A-rich sequence at their 3' ends. Many elements are truncated at their 5' ends. Complete elements have two long open reading frames. The first encodes a polypeptide like those in retroviral *gag* genes, and the other encodes a putative reverse transcriptase and RNase H (Jukubczak et al., 1990; Mathias et al., 1991).

I factor (I element)

I factor is the element that controls *I-R* hybrid dysgenesis in *Drosophila melanogaster*. Strains of *D. melanogaster* can be divided into two categories with respect to the *I-R* system: inducer strains which contain only complete and functional *I* factors, and reactive strains which contain only incomplete and non-functional *I* factors (Bucheton et al., 1976). *I* factors are stable in the genome of an inducer strain, but are active in the germ-line of female progeny of crosses between males of an inducer strain and females of a reactive strain. This results in reduced fertility and increased germ-line mutations. This phenomenon is known as *I-R* hybrid dysgenesis. It is restricted to female progeny as the male progeny of the same dysgenic cross are normal and show no increase in *I* factor transposition.

The *I* factor is 5.4 kb in size and has two open reading frames (Fawcett et al., 1986). The first encodes a nucleic acid binding protein with a cysteine-rich motif similar to that found in retroviral *gag* gene products (Dawson, et al., 1997). The second ORF encodes a protein with some of the motifs characteristic of a retroviral *pol* gene product and preliminary evidence indicates that it encodes a reverse transcriptase (Paterson, et al., unpublished data).

1.1.2 Class II elements and transposition

Class II elements are much more variable in organisation than class I elements. They range from simple insertion sequences (ISs), transposons, to relatively complicated bacteriophages. These elements all transpose directly from DNA to DNA without RNA intermediates. They normally have short inverted repeats bounding a single gene that encodes a transposase required for transposition. The transposons can be classified into different families according to the structure of the transposon, the structure of the encoded transposase, and the target site. These families are *P* family, hAT superfamily, *Tc1-Mariner* superfamily and *pogo* superfamily.

1.1.2.1 *P* family

The *P* element family, which was discovered to be responsible for *P-M* hybrid dysgenesis (Bregliano & Kidwell, 1983; Engels, 1983; Engels, 1989), is the best studied transposon in *Drosophila*. *P* element transposition is regulated genetically and occurs in a tissue-specific manner (Engels, 1983; Engels, et al., 1989). Molecular analysis of *P* element led to the development of the germline transformation method in *Drosophila* (Rubin & Spradling, 1982). *P* elements are now widely used as genetic tools and have revolutionised *Drosophila* molecular genetics.

The biological effects of *P* elements were initially observed in crosses between female of M strains, which lack functional *P* elements, and male of P strains, which carry active *P* elements. When a P-strain female is crossed to either a M-strain male or a P-strain male, the progeny are healthy, with a wild-type germ line. However. when an M-strain female is crossed to a P-strain male, the progeny exhibit a series of genetic disorders in the germline, such as chromosomal rearrangements, mutations, collectively referred to as hybrid dysgenesis. It was later discovered that the causative agent of hybrid dysgenesis is the *P*-element transposon (Bingham et al., 1982; Rubin, et al., 1982). These observations indicate that *P* elements introduced by P strain male into the egg of an M strain female mobilise in the germ line of the developing embryos and cause hybrid dysgenesis. In contrast, a P strain female maternally deposits *P*-element encode repressor proteins in her eggs that repress the mobilisation of any *P* elements introduced by the male. This state of repression in P strain females has been termed the P cytotype, while the state of permission in the M strain female has been termed the M cytotype (Engels, 1983; Engels, 1989).

P elements are heterogeneous and can be classified into two structurally distinct types: autonomous (full length) elements are 2902 bp in length, and non-autonomous elements that usually result from internal deletions of various sizes, and range from about 500-2500 bp (O'Hare & Rubin, 1983). In a typical P strain, there are 10-15 complete *P* elements and about 30-40 smaller elements derived from the complete elements by internal deletion. The complete 2.9 kb *P* elements contain four open reading frames (O'Hare & Rubin, 1983), which are transcribed to yield an mRNA of approximately 2.5 kb (Karess & Rubin, 1984). In somatic cells, splicing of the third intron is fully repressed, resulting in synthesis of the 66 kd protein that inhibits transposition (Misra & Rio, 1990; Rio et al., 1986; Robertoson & Engels, 1989). In the germline, spliced and unspliced mRNA are produced, yielding both the 87 kd transposase and the 66 kd repressor. This alternative splicing of the *P*-element transcript explains the germline-specific phenotypes associated with hybrid dysgenesis.

All the *P* elements analysed carry perfect inverted 31 bp terminal repeats, and create an 8 bp target-site duplication upon insertion. The transposase recognises and binds specifically to a 10 bp consensus sequence which is present at each end of the *P* element just inside the inverted repeats (Kaufman, et al., 1989; Mullins, et al., 1989). The host inverted repeat binding protein (IRBP) interacts specifically with the 31 bp inverted repeat (Beall, et al., 1994; Mullins, et al., 1989; Rio & Rubin, 1988). There is an internally located 11 bp inverted repeat that has been shown to act as a transcriptional enhancer *in vivo* (Mullins, et al., 1989) (Fig. 1.1-3).

P element derivatives with internal deletions can also encode proteins that repress transposition (Black, et al., 1987; Rasmusson et al., 1993; Simmons, et al., 1990). One of these repressor proteins is the 207 amino acid KP protein which has the first 199 and final 8 amino acids identical to the N-terminus and C-terminus of *P* transposase. Recent studies show that the KP protein binds to multiple sites at the ends of *P* element. These sites include the transposase binding site, the 11 bp transpositon enhancer, and the 31 bp inverted repeats. The DNA binding domain of KP protein is located at the N-terminal 98 amino acids which contains a CCHC metal binding motif. The KP protein can dimerise and contains two protein-protein interaction regions. This dimerisation is essential for high-affinity DNA binding (Lee, et al., 1996).

P elements are considered very recent invaders of *D.melanogster* because they are present in recently wild-caught strains, but not in lab stocks established in the first half of the century. This recent-invasion hypothesis is now generally accepted. *P* elements are not present in either the three sibling species or other species in the *melanogaster* species group, but are present in many other drosophilalids, especially the *willistoni* group (Daniels et al., 1990). The finding that a *P* element from *D.willistoni* was just a nucleotide different from that from *D. melanogaster* provided convincing evidence that *D. willistoni* or a sibling was the source of the *P* element that invaded *D.melanogaster*. *P* elements have also been found from several other *Drosophila* species including *D.nebulosa*, *D.bifasciata*, *D.guanche*, and *D. Subobscura*, and from another drosophilid fly *Scaptomyza pallida* (Lansman, et al., 1987; Paricio et al., 1992; Hagemann et al., 1992; Miller et al., 1992; Simonelig & Anxolabenhere, 1991). Phylogenetic comparisons indicate that two additional horizontal transfers must have occurred to produce the current distribution of these *P* elements in drosophlids (Clark, et al., 1994).

The molecular analysis of *P* elements led to the development of *P* element-mediated germline transformation (Rubin & Sparading, 1982). This procedure involves injection of embryos with two plasmid DNAs: one coding for transposase and the other carrying a genetic marker within a small, internally deleted *P* element.

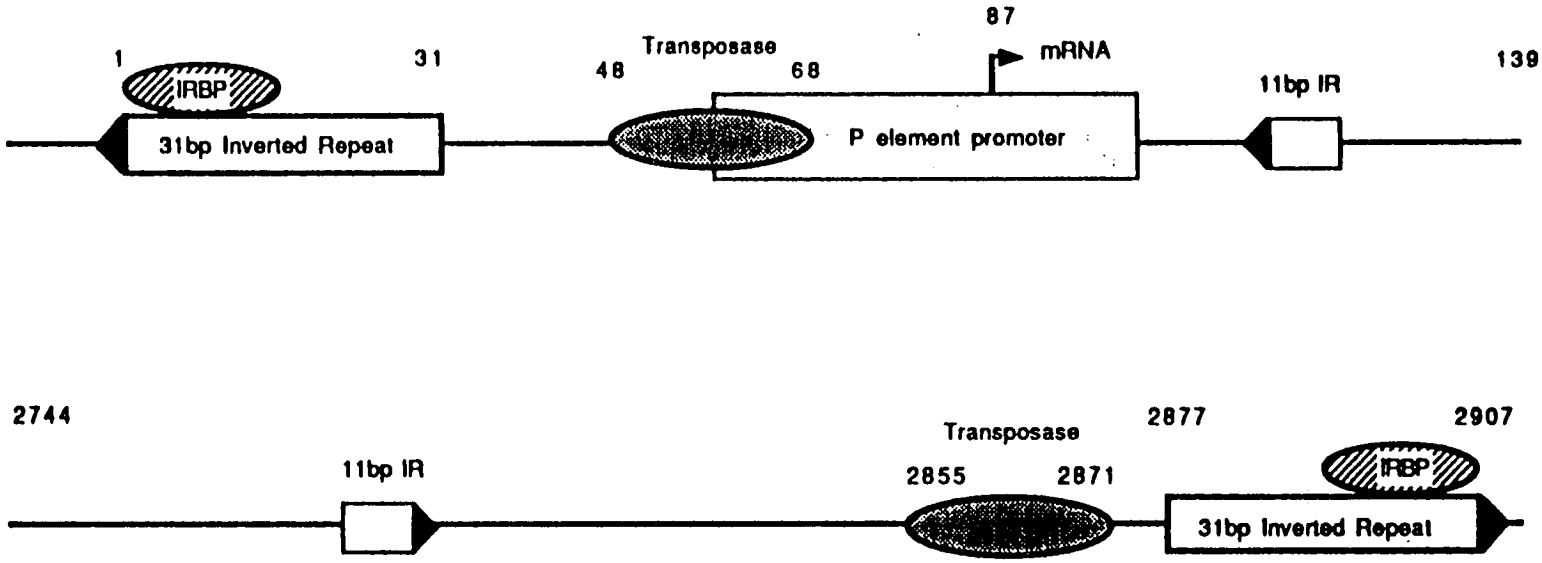


Fig. 1.1-3 The structure and DNA-protein interaction of the end seqences of a P element

After Rio, D. C. (1990)

The small nonautonomous *P* element contains the terminal cis-acting DNA sequences required for transposition, but does not encode transposase. Transposase, encoded by the first plasmid, catalyses transposition of the marked *P* element into the germline chromosomes. Following mating of the flies derived from the injected embryos, transformation is scored by the presence of the genetic marker within the marked *P* element. This procedure results in typical transformation frequency of 10-30 % (Spradling, 1986), and is now widely used.

The single *P* element insertion mutagenesis has revolutionised *Drosophila* molecular genetics (Cooley et al., 1988a&b; Berg & Spradling, 1991). It has later been usefully extended by the enhancer trap technique in which a reporter gene driven by a weak promoter is placed within the *P* element (O'Kane & Gehring, 1987; Bellen et al., 1989). The weak promoter is susceptible to the influence of adjacent genomic enhancers and the resulting pattern of reporter gene expression may reveal the pattern of supression the gene into which the *P* element has landed. A large number of single *P* element enhancer trap lines have been generated (O'Kane & Fehring, 1987; Bellen et al., 1989; Bier et al., 1989; Wilson et al., 1989). However, it is unlikely that a complete sample of the genome will be achieved using *P* elements alone, since genes display a wide variation in their receptivity to insertion (Kidwell, 1986; Engels, 1989). It is thought that about 50% of *Drosophila* genes are mutable by single *P* element mutagenesis (Cooley et al., 1988). This suggests that many genes will not be readily identified using *P* element mutagenesis alone. Moreover, *P* elements cannot be used as a transformation system for insect species other than *Drosophila melanogaster,* because the apparent dependence of the *P* element on a host factor in *D. melanogaster* restricts its mobility to closely related species. This may restrict their ability to transfer horizontally.

1.1.2.2 hAT superfamily

The hAT supefamily contains *hobo* elements from *Drosophila malanogster*, *Ac* elements from *Zea mays* and *Tam3* elements from *Antirrhinum majus* . The relationship among these three transposon families has been revealed by recent studies (Calvi et al., 1991; Feldmar & Kunze, et al.; 1991; Atkinson. et al., 1993). These three elements exhibit low levels of coding sequence similarity that span several hundred codons. All three elements generate 8 bp target site duplications upon insertion. They leave empty excision sites that are distinctly different from those observed for other transposable elements (Pohlman et al., 1984; Coen et al., 1986; Atkinson, et al., 1993). These data suggest that *hobo*, *Ac* and *Tam3* belong to a family of related transposable elements, named the hAT element family by Atkinson, et al.

(1993). Since the *Ac* and *Tam3* sequences appear more closely related to one another than either is to *hobo*, it has been proposed that these three elements have evolved from an ancestral element present prior to the divergence of plants and animals (Calvi et al., 1991). An alternation explanation for the conservation among these three transposon families is that it represents an example of horizontal transmission of genetic information between plants and animals.

hobo family

Hobo is a transposable element originally discovered in *Drosophila melanogaster* by its association with the *Sgs4* gene (McGinnis et al., 1983; Streck et al., 1986). The complete element HFL1 sequence is 2959 bp long which contains a 2.0 kb open reading frame (ORF1) encoding a putative transposase. *hobo* elements contain 12 bp terminal inverted repeats and create 8 bp target site duplications upon insertion. *hobo* excision results in complete removal of *hobo* sequences plus the addition of nucleotides at the resulting breaking points that appear to be simple or complex duplications of the flanking genomic DNA.

hobo elements are heterogeneous both in the number of copies in different *Drosophila* strains and in the size of different copies of the element within a single genome. Therefore *Drosophila* strains can be divided into two types: *hobo*-containing strains (H-strains) which include at least one complete hobo element, and empty strains (E-strains) which lack the complete element but contain a few *hobo* elements of smaller sizes. *hobo* has been shown to be able to act as inducers of *hobo* dysgenesis (H-E dysgenesis). It occurs mainly in the progeny of crosses between H males and E females. *hobo* elements can be used as effectively for enhancer trap mutagenesis as the *P* element (Smith et al., 1993). However, a *hobo* enhancer trap element has a pattern of insertion into the genome that is different from the distribution of *P* elements. The different insertion specificity make *hobo* insertion mutagenesis and enhancer tapping a valuable additional tool for more complete coverage of the *Drosophila* genome. In addition to its mobility in dysgenic crosses, the mobility of *hobo* also occurs in intrastrain crosses, producing molecular rearrangements such as inversions, deletions or new *hobo* insertions close to the resident element (Blackman et al., 1987; Yannopoulos et al., 1987; Lim, 1988). These rearrangements could be a consequence of recombination between two neighbouring *hobo* elements.

Like *P* elements, *hobo* elements are active predominantly in the germline of *Drosophila* (Blackman, et al., 1987; Blackman & Gelbart, 1989). However, this germline specificity is due to regulation of transposase production at the level of

transcription (Calvi & Gelbart, 1994) while the germline specificity of *P* element is regulated by RNA splicing (Laski et al., 1986; Rio et al., 1986).

hobo elements are present in all members of the *D. melanogaster* species groups, *D. melanogaster*, *D. simulans*, *D.mauritiana* and *D.sechellia*, and are widely distributed among species of two of its subgroups, *montrium* and *melanogaster* (Daniels, et al., 1990). Surveys of older isofemale lines of *D.melanogaster* suggest that complete *hobo* element were absent prior to 50 years ago. DNA sequences of *hobo* elements from *D.malanogaster*, *D. simulans*, and *D. mauritiana* reveal that *hobo* has recently been introduced into these species by horizontal transfer (Simmons, 1992).

Ac family

The maize Activator (*Ac*) element was discovered and characterised genetically by McClintock (1949, 1951,1965). It is 4565 bp long, has 11 bp inverted repeats and creates 8 bp target site duplication upon insertion (Muller-Neumann et al., 1984; Plhlman et al., 1984 a&b; Dooner et al., 1988). The five exons encode an 807 amino acid transposase protein which is the only *Ac* product required for transposition (Kunze et al., 1987; Li & Starlinger, 1990; Fusswinkel et al., 1991).

The transposase binds to several subterminal sequences of *Ac* but not the 11 bp long inverted repeats of the element. Major *Ac* transposase binding sites were delineated on 60 and 70 bp long sequence segments that lie 100 bp inside of the 5' *Ac* terminus and 40 bp inside of the 3' terminus respectively. The 5' end binding site overlaps the putative transposase gene promoter. Within all strongly bounded fragments, and particularly in these 60 or 70 bp long segments, the hexamer motif AAACGG is repeated several times in direct or inverted orientation (Kunze & Starling, 1989).

The transposase domain that recognises the subterminal DNA motifs is high in positively charged amino acids located around 200 aa from N-terminal. The N-terminal 102 amino acids of the transposase are not required for the transposition reaction (Feldmar & Kunze, 1991). Transposase mutants, deficient in DNA binding have been produced, which have *trans*-dominant inhibitor effect, suggesting that *Ac* transposase act as a dimer or multimer (Kunze et al., 1993). *Ac* transposase has been shown to be able to act as a transcriptional repressor to regulate the promoter activity.

The sequence of the inverted repeats of *Ac* is similar to sequences of the inverted repeats of transposable elements *Tam3* from *Antirrhinum majus* (Sommer et al., 1985), *hobo* and *P* from *Drosophilla* (O'Hare & Rubin, 1983; Streck et al., 1986).

Tam3 family

Tam3 element in *Antirrhinum majus* is 3629 bp long and contains an open reading frame of 2.48 kb as demonstrated by cDNA analysis. The single ORF encodes a predicted transposase of 748 amino acids. *Tam3* has the terminal inverted repeats similare to those of *Ac* and generates 8 bp target site duplication upon insertion. Moreover, patches of strong homology (60-65%) were found between the deduced *Tam3* protein and the amino acid sequence of the *Ac* gene product. The overall homology between both amino acid sequences is 30% over a stretch of 520 amino acids. The homologous part comprises the region between amino acids 220 to 740 of the *Ac* ORFA and 190 to 710 of the deduced *Tam3* protein (Hehl et al., 1991).

1.1.2.3 *Tc1-mariner* superfamily

One of the distinctive features of the *Tc1-mariner* superfamily of transposons is that they are widespread among different species: fungi, cilates, planarians, arthropods, nematodes, and vertebrates (Harris et al., 1988, 1990; Tausta & Klobutcher, 1989: Brezinsky et al., 1990; Franz & Savakis, 1991; Abad et al., 1991; Henikoff, 1992; Heierhorst et al., 1992; Robertson et al., 1992; Tudor et al., 1992; Daboussi et al., 1992; Robertson et al., 1993; Garcia-Fernandez et al., 1993; Caizzi et al., 1993; Radice et al., 1994). Both vertical and horizontal transfer have contributed to the spread of these elements (Robertson, 1993; Radice et al., 1994; Robertson and Lampe., 1995). The widespread occurrence of these transposons probably indicates that they are less dependent on species-specific host factors than other elements.

Elements in this family have the conserved sequence CAGT at their termini, and duplicate the TA dinucleotide upon insertion (Brezinsky et al., 1990; Franz & Savakis, 1991; Henikoff, 1992; Heierhorst et al., 1992; Tudor et al., 1992; Daboussi et al., 1992; Robertson et al., 1993; Radice et al., 1994). The transposase encoded by these elements have the conserved DDE catalytic domain shared with bacterial transposase and retroviral integrases (Robertson, 1993; Doak et al., 1994). The space between the second D and the E is 34 in this superfamily instead of 35 for prokaryotic transposons. The last residue of the DDE triad is D for *mariner* elements instead of the E for the others. This superfamily includes *Tc1* family and *Mariner* family.

Tc1 family

Tc1 and *Tc3* elements from *Caenorhabditis elegans* are examples of this family.

Tc1 element

Tc1 from *C. elegans* is an element of 1612 bp long, with short terminal inverted repeats and a large open reading frame. Although there is some variations in the exact DNA sequence (Rosenzweig, et al., 1983; Plasterk, 1987; Harris, et al., 1988), all *Tc1* elements have the same structure, no *Tc-1* elements of different size have been detected (Mori, 1988). The open reading frame encodes the 343 amino acids *Tc1* transposase, Tc1A (Emmons et al., 1983; Rosenzweig et al., 1983; Vos et al., 1993). The transposase binding site is located between base pair 5 and 26 from the end of *Tc1*, within the 54 bp inverted repeat. The N-terminal 63 amino acids of Tc1A bind specifically to the transposase binding site. A second, non-specific, DNA binding domain is located between amino acids 71 and 207 of Tc1A. Both the DNA binding domain of Tc1A and the DNA binding site in the inverted repeat of *Tc1* can be divided into two subdomains. The N-terminal part of the bipartite domain interacts with base pair 12-25 *via* the adjacent minor and major grove, while the C-terminal part of the bipartite domain shows additional contacts with bast pair 7-12 mainly *via* the major groove (Vos & Plasterk, 1994).

Tc1 was the first member of the *Tc1-mariner* family the transposase of which was shown to catalyse endonuclease cleavage and phosphate transfer reactions *in vitro* (Vos & Plasterk, 1994). Later studies show that the Tc1A transposase is the only nematode protein required for in vitro transpositon of *Tc1*. *Tc1* transposes *via* a cut and past mechanism. The terminal 26 bp sequence of the 54 bp terminal inverted repeats together with the flanking TA sequence are sufficient for this transposition. The target site choice *in vitro* is similar to that *in vivo*. This simple reaction requirements explain why horizontal spread of *Tc1/mariner* transposons can occur. They also suggest that *Tc1* may be a good vector for transgenesis of diverse animal species (Vos, 1996).

Tc1A has the conserved DD34E residues at its central region. A mutation in this DDE motif abolishes both endonuclease cleavage and phosphate transfer activities, suggesting that Tc1A carries a catalytic core common to retroviral integrases and IS transposases.(Vos & Plasterk, 1994)

Tc3 element

The *Tc3* element from *C.elegans* is 2335 bp long and has 462 bp inverted repeats at its ends. The two open reading frames of 417 bp and 570 bp respectively, encode the putative *Tc3* transposase (Van Luenen, et al., 1993).

The 65 amino acids at the N-terminal domain of *Tc3* transposase bind specifically to two conserved regions within the 462 bp *Tc3* inverted repeat; one region

is located at the end of the inverted repeat in between 11 bp and 30 bp, the other is located in between 184 bp and 202 bp from the end. The N-terminal DNA binding domain of the *Tc3* transposase interacts with nucleotides on one face of the DNA helix over adjacent major and minor groves (Colloms et al. 1994).

Analysis of the end sequences of the presumed transposition intermediate, the linear extrachromosomal DNA, shows that the *Tc3* transposon is excised incompletely. The 5' ends of the transposon lack two nucleotides, while the 3' ends coincide with the last nucleotide of the integrated element and carry 3' hydroxyls. This finding leads to the proposal of a model for *Tc3* transposition, which might apply to other transposon families transposing *via* the cut and past mechanism (van Luenen, et al., 1994) (Fig. 1.1-4). According to this model, transposition of *Tc3* is initiated by the binding of the *Tc3* transposase (and perhaps other cellular factors) to the ends of the transposon (Van Luenen et al., 1993). Double-strand breaks with a 2 bp stagger are generated, resulting in an excised element. The excised *Tc3* transposon contains the complete *Tc3* sequence at the 3' ends but lacks the terminal two nucleotides at each 5' end. The 3' end contains a 3' hydroxyl group, which could perform a nucleophilic attack during integration (Mizuuchi, 1992). Integration of *Tc3* results in a duplication of the TA target sequence. To generate this duplication, the 3' hydroxyl groups have to attack the phosphate bonds on both strands 5' of each T residue of the TA dinucleotide. After integration of *Tc3*, the 4 nucleotide gap at each end of the transposon will be repaired by the cellular machinery of the host to produce a complete *Tc3* element flanked by a duplicated TA sequence.

The target choice of the related transposable elements *Tc1* and *Tc3* was investigated by determining the exact location of 204 independent *Tc1* insertions and 166 *Tc3* insertions (Luenen & Plasterk, 1994). All insertions were shown to be into the sequence TA. The integration sites did not reveal a strong consensus sequence for either transposon.

Mariner family

This family of transposons is characterised by relatively small size of approximately 1300 bp, short inverted terminal repeats of about 30 bp, the induction of a TA dinucleotide target-site duplication upon insertion, and the presence of a single open reading frame (ORF) presumed to encode a transposase (Jacobson et al., 1986; Medhora et al., 1991). This is a very diverse family of transposons, all the elements in this family have been called *mariners* or *mariner*-like elements.

The *mariner* element was first detected by its association with the mutation *white peach* of the white eye gene (*w*) in *D. mauritiana*. Frequent somatic excision in

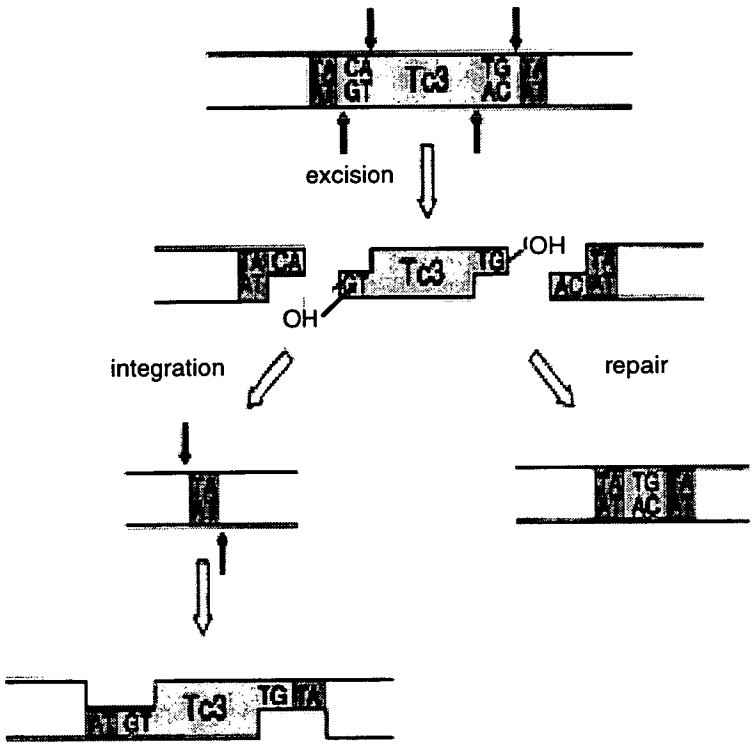


Fig. 1.1-4 Model for Tc3 transposition

(After van Luenen et al., 1994)

one strain of *D. mauritiana* caused by excision factor *mos* had been observed (Bryan et al., 1987). The *mos* factor, which is a particular copy of *mariner*, is believed to activate transposition of itself and other *mariner* elements by being able to supply transposase in *trans*. Germline excisions due to the activity of *mariner* have also been reported (Haymer & Marsh, 1986; Jacobson & Hartl, 1985). The *mariner* element is 1286 bp long and contains 28 bp terminal inverted repeats. It creates TA dinuclotide target site duplication upon insertion. The 1035 bp single open reading frame was predicted to encode a 345 amino acid transposase (Hartl, 1989).

Almost all copies of *mariner* studied so far have been uniform in length (Hartl, 1989), they differ from each other generally by single nucleotide changes (Maruyama et al., 1991). Non-autonomous elements are not obvious deletions of autonomous elements. Site directed mutagenesis within the putative transposase coding region appears to be responsible for inactivation supporting the hypothesis that the ORF encodes a transposase (Maruyama et al., 1991). The activity level of elements is also greatly influenced by immediately adjacent sequences (Medhora et al., 1991).

mariner occur in five of eight species of the *D. melanogaster* species subgroup, *D. yakuba*, *D. teissieri*, *D. simulans*, *D. mauritiana* and *D. Sechellia*, but not in *D.melanogaster* (Lachaise et al., 1988; Jacobson et al., 1986; Maruyama et al., 1991; Capy et al., 1992). However, *mariner* sequences have been detected in more distantly related drosophilids (Maruyama & Hartl, 1991), in ten other insects representing six additional orders (Robertson, 1993), in lower invertebrate planarian (Garcia-Femandez et al., 1993), in nematodes and flatworms (Robertson, 1995; Robertson & Asplund, 1996), and also in human and other mammalian genomes (Simt & Riggs, 1996; Augegouillou, et al., 1995; Morgan, 1995; Oosumi, et al., 1995). The DNA sequences of these elements are much more conserved than the rest of the genome, which strongly indicates that horizontal transfer is responsible for the wild distribution of *mariner* element.

The extremely widespread distribution of *mariner* elements leads to the hypothesis that *mariner* transposition relies solely on the *mariner* transposase and not species-specific host factors beyond those necessary to transcribe and translate the transposase and host repair enzymes necessary to repair single-stranded gap at the site of transposition insertions (Lampe et al., 1996). This hypothesis is strongly supported by recent experiment which show a purified *mariner* transposase is sufficient to mediate transposition *in vitro* (Lampe et al., 1996). In this experiment, the transposase of a *mariner* (*Himar 1*) from the horn fly, *Haematobia irritans*, was purified. It was shown to be sufficient to reproduce transposition in an *in vitro* inter-plasmid transposition reaction. This transposase was also shown to be able to bind to the

inverted terminal repeat sequences of the transposon and mediate 5' and 3' cleavage of the element termini. The later indicates that *mariner* transposes by a cut and paste mechanism similar to that proposed for *Tc3* transposition (van Leunen et al., 1994).

1.1.2.4 *pogo* superfamily

This superfamily is named after the original *pogo* element of *Drosophila melanogaster*. All the transposons in this family share significant amino acid sequence similarity with that of *pogo* in the central conserved and alignable DDE region. Like members of the *Tc1-mariner* superfamily, *pogo* elements create TA dinucleotides target site duplication upon insertion and have short terminal inverted repeats at both ends. The major difference between these two families is that there is no proper candidate for the last E/D residue in the DDE motif for the *pogo* family transposons (Doak, 1994; Robertson, 1996). It is this character that distinguishes the *pogo* superfamily from the closely related *Tc1-Mariner* superfamily. Transposons in the *pogo* superfamily can be divided into three families: the *pogo* family, the *Fot1* family and the *Tc4* family. Moreover, the transposases encoded by members of the *pogo* superfamily can also find significant similarities with some nontransposase proteins, such as mammalian centromere protein CENP-B, the *jerky* protein which is involved in epileptic seizures in mice, the yeast regulatory proteins RAG3 and PDC2. However, the significance underneath these similarities still need to be further investigated.

pogo family

This family contains *pogo* element from *Drosophila malanogaster* (Tudor et al., 1992) and *Tiggers*, the *pogo* transposon fossils in the human genome (Smit & Riggs, 1996).

pogo elements

The *pogo* element was first discovered as a 190 bp insertion associated with *white-eosin* (w^e) mutation in *Drosophila melanogaster*. A *Doc* element which inserted in the promoter region of the *white* (*w*) locus has been shown to be responsible for the w^1 mutation (Driver et al. 1989). The 190 bp *pogo* element, as an additional secondary insertion within the *Doc* element, is responsible for the w^e mutation (Tudor et al. 1992).

This 190 bp *pogo* element was used as a probe in DNA blotting to examinate the sizes of *pogo* elements in three differernt wild type strains, Canton S, Oregon R and Harwich. The result shows that strains typically have many copies of the 190 bp element, 10-15 copies of the 1.1-1.5 kb element, and several copies of the 2.1 kb element. The smaller elements all appear to be derived from the largest by single

internal deletions so that all elements share the same terminal sequences (Tudor, et al., 1992).

The insertion sites of *pogo* elements are generally AT rich. *pogo* elements always insert into the target site TA and make a 2 bp duplication upon insertion. This is similar to the *Tc1* element of *C. elegans* (Moerman and Waterson, 1989). The terminal 4 bp CAGT sequence of *pogo* is also identical to that of *Tc1*.

The DNA sequence of a full length of *pogo* element is 2121 bp with 21 bp terminal inverted repeats, and has two open reading frames of 1155 bp and 417 bp respectively (Fig. 1.1-5). cDNA analysis indicates that the two ORFs are joined by RNA splicing to encode for a single polypeptide which is supposed to be the putative *pogo* transposase. Compared to standard codon usage in *D.melanogaster*, codons rich in A or T are preferred in *pogo* (Tudor, et al., 1992).

The predicted protein sequence of the *pogo* transposase encoded by ORF1-ORF2 has significant homology with human centromere protein B (CENP-B) (Earnshaw et al. 1987). The proteins show 51% similarity and 26% identity and can be aligned for almost their entire lengths. Another similarity is to the *jerky* protein of mice, which is involved in epileptic seizures. The proteins have 52.9% similarity and 26.3% identity over the complete sequences (Toth et al. 1995). *pogo* transposase can also find significant similarities to two yeast transcriptional factor proteins, PDC2 and RAG3.

The various copies of *pogo* within the *D. malanogaster* genome differ by internal deletions and less than 1% sequence divergence, indicating that it has been in this genome for a relatively short time. Examination of the genomic occurrence of *pogo* elements in over 120 strains of *D. melanogaster* has shown that there are differences between strains in their intensities or presence of deletion-derivative of *pogo*, suggesting that these elements are currently active or have been so recently. *pogo* elements have not been detected in eleven other species in the genus including the sibling species *D. simulans* and *D. mauritiana*, three different subgenera, and the related *Scaptomyza pallida*. These results suggest that *pogo* element may have horizontally transferred into *D. malanogaster* since its divergence from its sibling species (Boussy, et al., 1993).

Tigger elements

Tiggers, which were found by analysis of published fragments of human medium reiterated frequency repeats (MERs), are interspersed repeats which resemble fossils of DNA transposons by excision and integration into the genome without an RNA intermediate. The 2417 bp consensus sequence of *Tigger1* contains two long

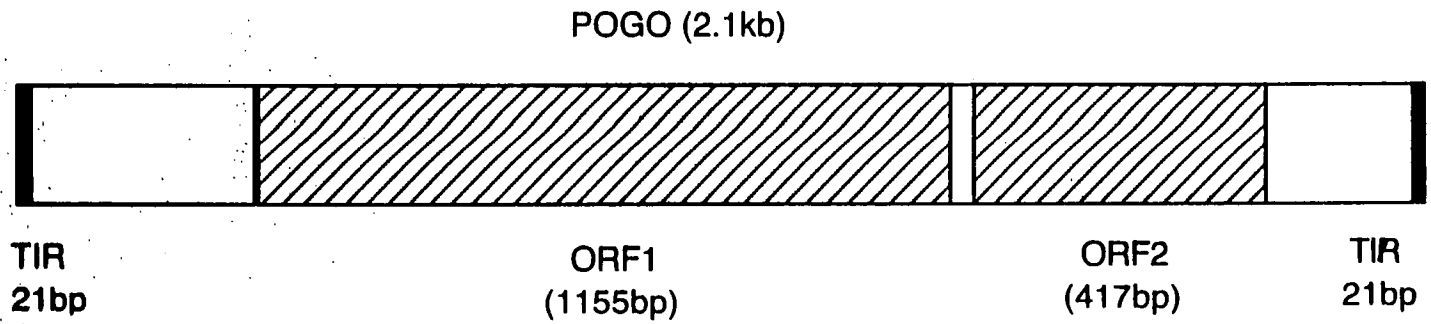


Fig. 1.1-5 Structure of transposable element pogo

open reading frames, one of which is 1335 bp and encodes a product closely related to the putative transposase of *pogo* element. *Tigger2* has a 2708 bp consensus sequence. The products of *Tigger1* and *Tigger2* are 48% identical, whereas their DNA sequences, aligned as guided by their products, are only 54% similar in the coding region. The internally deleted *Tigger2* products are much more common than the full length one. This pattern is very similar to that of *pogo*, which has many copies of a 190 bp internal deletion product, 10-15 copies of an approximately 1.3 kb element, and only a few full length *pogo* sequence. In contrast, *Tigger1* seems primarily represented by full length elements, only a couple of internal deletion products were found (Smit & Riggs, 1996).

Fot1 family

This family contains three elements from fungi with their transposase sequences most similar to that of *pogo* in the DDE region. These elements are the *Fot1* element of *Fusarium oxysporum* (Daboussi et al. 1992), the *Pot2* element of *Magnaporthe grisea* (Kachroo et al. 1994), and the *Fcc1* element of *Cochliobolus carbonum* (Panaccione et al.1996). The transposase of these elements are approximately the same length as that of *pogo* and they share 21-26% amino acid similarity in the central conserved and alignable DDE region.

Fot1 elements

Fot1 is a family of transposable elements found in the fungal plant pathogen *Fusarium oxysporum*. The first element was identified as an insertion in the gene encoding nitrate reductase. It is 1928 bp long, has 44 bp terminal inverted repeats, and produce TA dinucleotide target site duplication upon insertion. It has a single 1626 bp open reading frame potentially encoding a 542 amino acid polypeptide, the transposase (Daboussi et al, 1992).

Pot2 elements

Pot2 is a transposable element from *Magmaporthe grisea.* The element is 1857 bp long with 43 bp terminal inverted repeats (TIR) and 16 bp direct repeats within the TIR. It has an open reading frame of 1605 bp, potentially encoding a polypeptide of 535 amino acids. Two additional in-frame ORFs potentially coding for polypeptides of size 374 and 252, respectively, were identified within the larger ORF.

The putative polypeptide encoded by the lagest ORF of *Pot2* showed extensive identity to that of *Fot1*. The two polypeptides, which are different in length by seven amino acids, have 39.3% identity, and varying levels of homology are present

throughout the entire length of the protein, with the highest conservation being in the middle region (amino acid 160-380) of the protein which includes the conserved DDE region.

Pot2 also produce TA dinucleotide duplication upon insertion. The consensus insertion sequence of *Pot2* revealed certain feature similarity to that of *Tc1* element. *Pot2* is present at a copy number of approximately 100 per haploid genome and represents one of the major repetitive DNAs shared by both rice and non-rice pathogens of *M.grisea.*

Fcc1 elements

Fcc1 (*Fot1*-like element from *Cochliobolus carbonum.*) is a repetitive element found in fungus *C. carbonum.* The two potential TIRs of this putative transposon are most similar to each other and have structural similarities to transposons of the *Fot1* family. *Fcc1* also has a similar size and structure to *Fot1* and *Pot2* elements. Its DNA sequence is 49.5% identical to that of *Fot1* element over 1660 bp, and 44% identical to that of *Pot2* element. The amino acid sequence of *Fcc1* is 27 % identical to that of *Fot1* and 24 % identical to that of *Pot2.*

The *Fcc1* element was found in varying copy number in the genomes of all *Cochliobolus spp.* examined, giving a distinct fingerprint in each species and race tested (Panaccione et al., 1996).

Tc4 family

This family contains three transposons from the nematode *C. elegans* which are more distantly related to *pogo.* The *Tc4* (Yuan et al. 1991; Li and Shaw 1993) and *Tc5* (Collins and Anderson 1994) elements encode transposases with 19% and 14% identity to the D,D35E region of *pogo* respectively. The *Tc2* element (Ruvolo et al.1992) encodes a polypeptid having similarity to the region surrounding the first D of *pogo* transposase, no information about the comparison of the rest regions.

Tc4 elements

Tc4 is a distinctive family of transposable elements in the nematode *C. elegans.* The longest element found, named *Tc4v*, is 3483 bp in length, the others are 1605 bp long and consist of two 774 bp nearly perfect inverted repeats connected by 57 bp loop, and lacks significant open reading frames. In *Tc4v*, a 2343 bp novel sequence is present in place of a 477 bp segment in one of the inverted repeats. The novel sequence of *Tc4v* is present about five times per haploid genome and is invariably associated with *Tc4* element. cDNA analysis suggests that the element contains at least

five exons that could encode a novel basic protein of 537 amino acid residues (Li & Shaw, 1993).

Tc5 elements

Tc5 is a family of transposable element in the nematode *C. elegans*. *Tc5* is active in the *mut-2* strains. There are 4-6 copies of *Tc5* per haploid genome. *Tc5* is a 3.2 kb long element bounded by inverted terminal repeats of nearly 500 bp. Individual *Tc5* elements are similar to each other in size and structure. *Tc5* elements are similar to *Tc4* elements in that eight of the ten terminal nucleotides of *Tc5* are identical to the corresponding nucleotides of *Tc4*, both elements recognise the same target site for insertion (CTNAG) and both cause duplication of the central TNA trinucleotide upon insertion. However, *Tc5* is unrelated to the three other transposon families from *C. elegans*, *Tc1*, *Tc2* and *Tc3*, which are also active in *mut-2* mutant strains (Collins & Anderson, 1994).

Tc2 elements

The *Tc2* element from *C. elegans* is 2074 bp in length and has inverted terminal repeats of 24 bp. There are three large open reading frames on one strand, which might encode the putative transposase protein. *Tc2* has unusual subterminal degenerate direct repeats that are arranged in an irregular overlapping pattern. Both of the two insertions examined create a 2 bp target site duplication on insertion (Ruvolo et al., 1992).

Proteins related to pogo-like transposases

CENP-B protein

Human centromere protein B (CENP-B) is a 80 kD protein which binds specifically to a 17 bp motif, CENP-B box, in centromeric alpha-satellite DNA. The DNA binding domain of CENP-B is located within the N-terminal 125 amino acid residues which contain four potential alpha-helices. CENP-B forms a stable complex, complex A, containing two alphoid DNAs *in vitro* (Masumoto, et al., 1989; Muro, et al., 1992; Yoda, et al., 1992). Polypeptides bind a DNA molecule as a monomer at the N-terminal DNA-binding domain, while dimerise at a novel hydrophobic domain in the C-terminal region of 59 amino acid residues. CENP-B polypeptides form a homodimer at the C-terminal hydrophobic domains, each binding a DNA strand at their N-terminal domains. The dimerisation and DNA binding domains fall into two of the three completely conserved sequences found in human and mouse CENP-B.

CENP-B is proposed to play an important role in the assembly of specific centromere structures by forming unique DNA-protein complexes at the sites of CENP-B boxes on the centromeric repetitive DNA both in interphase nuclei and on mitotic chromosomes (Kitagwa, et al., 1995).

jerky protein

jerky is a gene of 2148 bp long and has an open reading frame encoding a putative 41.7 kD protein of 370 amino acid residues in a single exon. The inactivation of this gene results in the absence or reduced level of the *jerky* protein and cause epileptic seizures in mice. The *jerky* protein has seven domains with significant homology to sequences represented in centromere protein-B (CENP-B), *Drosophila pogo*-R11 transposable element, and yeast regulatory proteins RAG3 and PDC2. The sequence alignments were more significant with CENP-B and *pogo*-R11. Overall, *jerky* showed 28% identity and 49.7% similarity to CENP-B, and 26.3% identity and 52.9% similarity to *pogo*-R11. These homologies suggest that *jerky* might function as a DNA binding protein. Of the four alpha helices in the N-terminal 125 amino acid DNA binding domain of CENP-B, three are also predicted in the *jerky* protein. In addition, *jerky* contains two more regions with potential alpha helical structure (Toth, et al., 1995).

PDC2 and RAG3 proteins

PDC2 and RAG3 are yeast transcription factors that are not associated with transposons. However the proteins encoded by these genes have been shown to have significant similarity to the transposases encoded by the transposons of the *pogo* superfamily.

RAG3 is a 2936 bp long gene in yeast *Kluyveromyces lactis*. It is a regulatory gene of pyruvate decarboxylase synthesis. It is required for the transcriptional regulation of the glucose transporter gene, RAG1 (GenBank accession no. X70186).

PDC2 is a transcriptional activator necessary for high level expression of pyruvate decarboxylase in *Saccaromyces cerevisiae*. The DNA sequence is 3184 bp long which encodes a predicted protein of 925 amino acids. The protein sequence is rich in asparagine and serine residues, which is often found for transcriptional regulators (Hohmann, 1993).

1.2 Unity in transposition reactions

Although different intermediates (DNA or RNA) are used in the transposition processes catalysed by transposase and integrase, the central reactions in them are similar. They all occur by similar DNA breakage and joining reactions, which are chemically identical. Moreover, the few available structures of transposases and integrases have remarkable overall structural similarity, even though they lack extensive primary sequence homology. They all have the similar sequence at the presumed catalytic domain, the common DDE motif, which has been show to be responsible for coordinating the essential divalent ion Mg^{2+} in the transposition reaction.

1.2.1 Transposition reactions

1.2.1.1 Transposition reactions catalysed by retroviral integrases

In a retrotransposition process, the integrase is responsible for catalysing the following three reactions (Craigie et al., 1990; Katz et al., 1990; Chow et al., 1992):

1. Processing of the double strand DNA ends, involving single-strand cleavage and removal of the terminal two 3' bases;

2. Integration of the cleaved DNA into the host chromosome, by a nucleophilic attack by the liberated 3' OH on the target DNA, leading to a single strand join .

3. Excision of the integrated oligonucleotide in a reaction called disintegration.

1.2.1.2 Transposition reactions catalysed by transposases

In the process of transposition, transposases are responsible for catalysing the following three reactions (Polard & Chandler, 1995):

1. Formation of a synaptic complex involving both transposon ends and the transposase;

2. A 3' strand cleavage at each end of the transposon to generate a 3' OH group;

3. Transfer of the cleaved transposon strands into a suitable target DNA, in which the target backbone phosphate group undergoes concerted nucleophilic attack by the 3' OH group on the cleaved donor strand.

1.2.2 The central reactions in transposition and integration

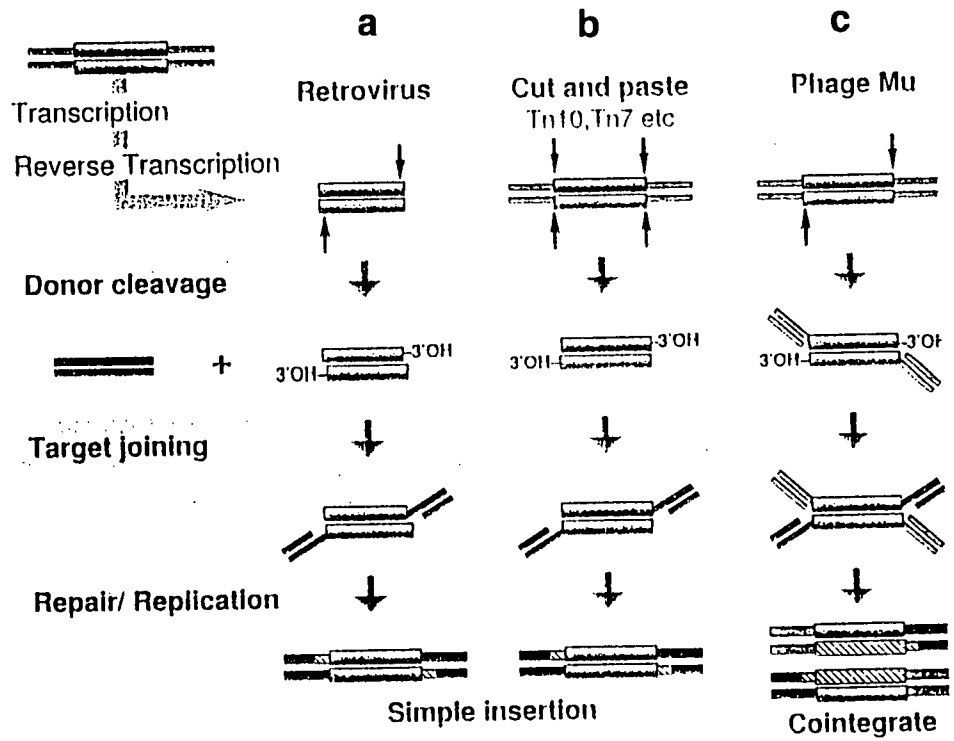
1.2.2.1 The central reactions in integration and transposition are similar

From the reactions catalysed by integrases and transposases described in 1.2.1, it is clear that central to these different processes, are the similar reactions of DNA cleavage and strand transfer. DNA cleavage reaction precisely exposes the 3' ends of both strands of the elements; and the strand transfer reaction joins the exposed 3' ends to the target DNA. Sometimes DNA cleavage reactions also expose the 5' ends of both strands of the elements, which will result in a transposition product of different structure. When both the 3' and 5' ends of the transposon are broken, the element is excised from the donor site and inserted into the target site to form directly a simple insertion through a cut and paste mechanism. When only 3' end cleavage occurs, the transposon remains attached to the donor site and is also attached to the target site. Processing of this strand transfer product by host replication functions generates a structure called a co-integrate in which two copies of the transposon fuse the donor backbone and the target DNAs. These similarities and differences are well represented in the processes of retrovirus integration, phage *Mu* transposition and *Tn10/Tn7* transposition described later (Fig.1.2-1 a-c).

DNA cleavage reactions

In this step, the transposon donor DNA is specifically cleaved at the two ends of the transposon to generate 3'-OH residuces. For some elements, such as phage *Mu* and retroviruses, the cleavage only occurs at the 3' ends, leaving the other strand uncleaved (Fig.1.2-1c). Other elements, such as *Tn7* and *Tn10*, which transpose *via* a cut and past mechanism, generate double strand cuts at the ends of the element (Fig.1.2-1b). For all the elements studied, the 3'-OH termini of the element exposed by the endonuclease cleavage is utilised for the next strand transfer step.

Retroelements also use the same DNA cleavage mechanism to expose the 3' ends of the elements. They transpose via an RNA intermediate that is converted into double strand linear DNA, which functions as the direct precursor for integration into a new location in the host chromosome. This integration process is similar to other transposition reactions. Most retroviral integration precursor DNAs are two base pairs longer at each end than the sequence that integrates into host chromosome. These extra nucleotides at the 3' end of each strand are removed by DNA cleavage (Fig.1.2-1a).



1.2-1 Transposition reactions of retrovirus, Tn10/Tn7 and phage Mu

After Craig (1995) and Polard & Chanddler (1995)

Strand transfer reactions

Strand transfer is the cleavage and joining of phosphodiester bonds that results in connection of the donor and target DNA. In this process, under the attack of the 3' OH group of the cleaved DNA, the target DNA is staggered cleaved to generate a 5' end extension. This is coupled with a single strand joining of the processed 3' hydroxyl ends of the transposon to the 5' phosphate ends of the target DNA to generate the target site duplication. The target DNA cleavage and joining to the donor DNA ends are energetically coupled so that no external energy sources such as ATP are required. All the elements studied so far join the 5' ends of the cut target DNA to the 3' OH ends of donor DNA, leaving the 3' ends of the target strands unjoined to the donor. These gaps are later joined by the host DNA repaire system.

The structure of the strand transfer product, depends on the nature of the unjoined donor DNA ends. For elements such as *Tn7* and *Tn10* that transpose by the cut and paste mechanism, gap repair completes the transposition process. The strand transfer product of phage *Mu* carries a pair of branched DNA structure, due to the remaining connection to the original donor DNA flanking the 5' ends of the transposon. Such an intermediate can be processed by replicating the entire transposon sequence and the short target site sequence.

1.2.2.2 The DNA cleavage and strand transfer reactions are chemically identical

The DNA cleavage and strand transfer reactions, which have been shown to be the central reactions of all transposition processes, are themselves chemically identical reactions. Both of them are Mg^{2+} dependent reactions and likely result from direct, one-step transesterifications without covalent protein-DNA intermediates (Mizuuchi, 1992). In DNA cleavage, H_2O is the nucleophile that attacks a phosphodiester bond hydrolysing and exposing of the 3' -OH transposon end. In strand transfer, the newly exposed 3'-OH transposon end is the attacking nucleophile that attacks by the 3' -OH transposon end on the target DNA results in the covalent linkage of the transposon end to a target strand. This chemical similarity suggests that the same, or at least closely related, active sites in the integrases and transposases execute both of these reactions.

1.2.3 The DDE common regions of transposases and integrases

Extensive sequence comparisons over the last few years have indicated that the similarities in mechanisms of integration and transposition are reflected in similar amino acid sequences of integrases and transposases. These similarities are confined chiefly to the catalytic domains of these proteins. The most notable feature of these

conserved regions is a triad of "invariant" carboxylate residues, two aspartates and a glutamate known as the DDE motif. These three conserved residues are well separated in the primary sequence with spacers of about 50-70 residues between the two aspartates and 34/35 residues between the second aspartate and the glutamate. The space between the latter D and E is usually 34 amino acids in the eukaryotic elements, but 35 in the prokaryotic elements (Doak et al., 1994; Polard and Chandler, 1995). This DDE motif has been found to be conserved in many integrase and transposase sequences encoded by completely different transposable elements. Some transposable element families can therefore be linked with the others by the DDE motif.

1.2.3.1 The DDE motif of integrases

The highly conserved DDE motif was first found in the integrase sequences of retrovirus and retrotransposons (Kulkosky, et al 1992; Fayet, et al., 1990; Johnson, et al.,1986; Rowlan & Dyke, 1990). The highlighted functional domains of 80 retroviruses and retrotransposons are highly conserved 50 amino acid sequence, which were referred to as D(35)E motif, since it is composed of invariant aspartate (D) and glutamate (E) residues consistently separated by 35 amino acids. The residues in this region were supposed to contribute to the common functions of DNA recognition, cutting, and joining shared by these proteins (Kulkosky, et al., 1992).

1.2.3.2 The DDE motif of transposases

The highly conserved DDE motif has been found in more and more families of the class II elements including many bacterial insertion sequences (ISs), bacterial transposon *Tn7*, elements of the *Tc1-mariner* superfamily, hAT superfamily, *pogo* superfamily, and bacterial phage *Mu* (Table 1.2-1). This motif is also supposed to be essential for transposase activity. So transposases are linked with integrases by the common DDE motif, and so do the transposon families with that of retrovirus and retrotransposons.

The DDE motif of bacterial insertion sequences (ISs)

Members of the *IS3* insertion sequence family were the first class II elements found to have the conserved DDE motif (Kulkosky, et al 1992; Fayet, et al., 1990;). The DDE motif in the *IS3* family is similar to the D(35)E motif in retroelements. Later studies show that this motif is also present in the transposases of the members of the *IS630* family in *Shigella* (Henikoff, 1992)

Table 1.2-1 The DDE motifs of some transposases

Transposon	DDE motif	comments
IS3 & IS360	D, D(35)E	prokaryotic transposon; same DD(35)E motif as retroelements.
Tn7	D, D(35)E	prokaryotic transposon; same DD(35)E motif as retroelements.
Tc1 & Tc3	D, D(34)E	eukaryotic transposon; last D and E seperated by 34 residues.
Mariner	D, D(34)D	eukaryotic transposon; E is subsititued by D; last two D seperated by 34 residues.
hAT elements	D,S/D,E	second D is substituted by Serine(S) in Ac and some hobo transposases; space between second S/D and last E varies from 34 to 37.
pogo elements	D,D,?	no proper candidate for the last residue of the DDE trait.
phage Mu	D,D,E	second D and E separated by 56 residues.

The DDE motif of bacterial transposon *Tn7*

The TnsA and TnsB subunits of the transposase of bacterial transposon *Tn7* have both been shown to contain the conserved DDE motif. Like other prokaryotic elements, the last D and the E are generally separated by 35 amino acid residues (Radstrom et al., 1994; Craig, 1996c; Polard & Chandler, 1995; Sarnovsky et al., 1996).

The DDE motif of *Tc1-mariner* superfamily

Members of the *Tc1-mariner* superfamily are the eukaryotic elements found having the conserved DDE motif. The space between the latter D and E is 34 here, rather than 35 in prokaryotic elements. All three residues of the D,D(34)E motif are present in the transposase coded by *Tc1* elements, while the distal E has undergone a chemically conservative E-D substitution in the *mariner* elements (Doak et al., 1994).

The DDE motif of hAT superfamily

Comparison of the transposases coded by hAT elements and *Tc1-mariner* elements shows that the DDE motif may also be present in hAT sequences (Bigot, et al., 1996). The second amino acid of the DDE triad is replaced by a serine in the *Ac*, HFL1 *hobo* and *hermes* transposases. There are also variations in the number of residues between the first and the second acidic residues of the putative DDE motif in all these transposases. But the alignment shows that the differences in length are due to insertions or deletions at the same position.

The DDE motif of *pogo* family

The putative *pogo* transposase sequence has a sequence similar to the DDE motif, but the final conserved glutamic acid is not present (Doak et al., 1994; Robertson, 1996). All the other members of the *pogo* superfamily have significant amino acid sequences similar to that of *pogo* in the central conserved DDE region, suggesting that they might be related to the extended DDE superfamily of transposase and integrase (Robertson, 1996).

The DDE motif of phage *Mu*

The transposase of phage *Mu,* MuA, has also been show having a DDE motif. Amino acid sequence alignment first suggests that Asp-269 and Glu-392 of MuA may correspond to the first Asp and last Glu in the DDE motif, but there is no obvious candidate for the middle Asp (Baker and Luo, 1994). Later studies (Kim. et al., 1995) show that Asp-294 could be the middle D in the DDE motif. However, structure

studies reveals that Asp-336, together with Asp-269 and Glu-392, make up the DDE motif of MuA transposase (Grindley & Leschziner, 1995).

1.2.4 The role of DDE motif in transposition reactions

It has been shown in 1.2.2 that the reactions catalysed by integrases and transposases are similar. DNA cleavage and strand transfer are identical Mg^{2+} dependent chemical reactions. And it has been show in 1.2.3 that integrases and transposases are structurally similar as they all have the conserved DDE motif in their catalytic domains, which have been show to be essential for catalysing transposition reactions. So it is very likely that the DDE motif is closely related to the central reactions of transposition. This has been investigated by changing these residues by site directed mutagenesis and by determining the three dimentional structures of some of these proteins. The results from both studies indicate that the DDE motif forms part of the active site of these enzymes and probably contributes to their activity by coordinating divalent metal ion involved in catalysis (Baker & Lou, 1994; May & Craig, 1996; Balland & Kleckner, 1996; Katz & Skalka, 1994).

1.2.4.1 Mutagenesis studies

Mutagenesis studies were conducted to change amino acid residues in the catalytic domain of the integrases and transposases, to analyse their effects on the transposition. The results show that replacement of the conserved D or E residues in the DDE region dramatically reduced the 3' end cleavage and joining activities, even when the substitutions were highly conservative (Engelman & Craigie, 1992; Kulkosky, et al., 1992; van Gent, et al., 1992; Baker & Luo, 1994; Kim et al., 1995). The dependence of both cleavage and joining on these invariant residues suggests that they participate in both reactions. Since acidic residues are characteristically involved in metal binding, and both the cleavage and joining reactions of integrase/transposase require Mg^{2+} or Mn^{2+}, one possible role for these invariant residues would be coordination of the metal cofactor(s) required in these reactions.

1.2.4.2 Structure analysis

More direct evidence concerning the role of the DDE motif in catalysing transposition reactions comes from structural analysis. The X-ray crystal structures of the catalytic core domain of HIV integrase, *Mu* transposase and ASV integrase have been analysed, and provide more information about the DDE catalytic motif (Fig. 1.2-2a-c).

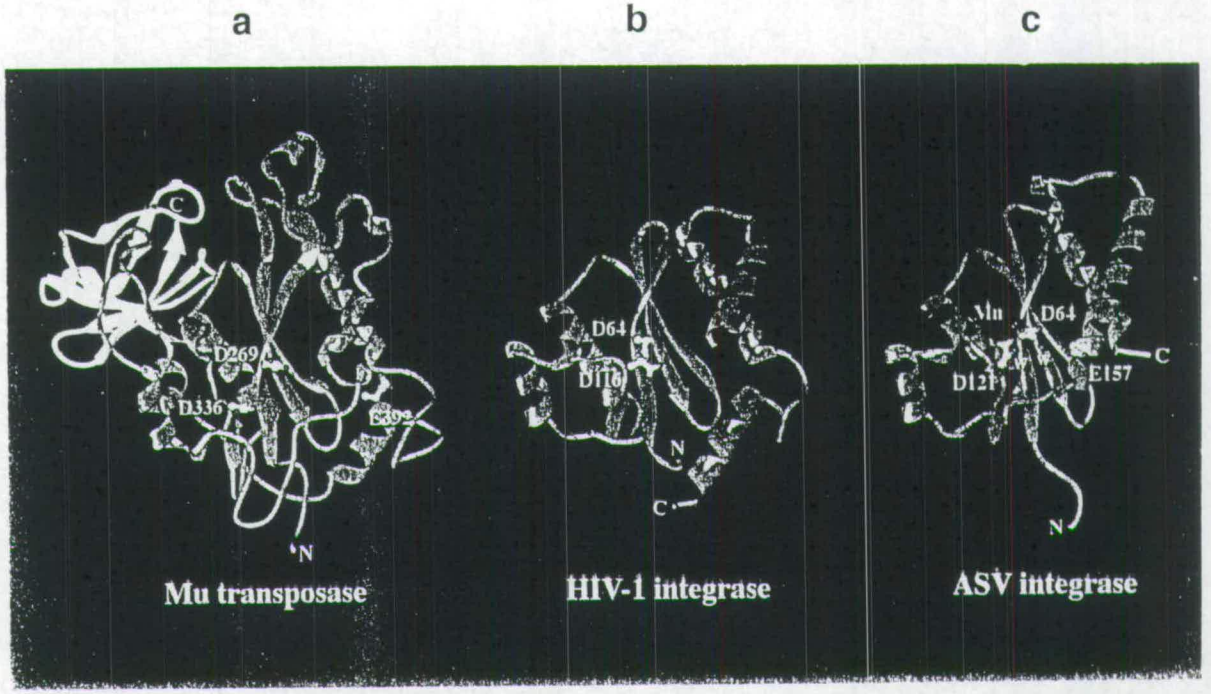


Fig. 1.2-2 Structure of the catalytic domains of Mu transposase, HIV integrase, and ASV integrase

(After Grindly & Andres, 1995)

The structure of HIV integrase

In the analysed structure of HIV integrase, only the two aspartate residues (Asp-64 and Asp-116) of the DDE triad were visible. They were close to one another, however, and the third carboxylate, Glu-152, located on a 13 residue disordered region, seemed likely to be located in the same general area. This result is compatible with the proposed role of the three carboxylates in metal binding and catalysis (Dyda et al., 1994) (Fig. 1.2-2b).

The structure of *Mu* transposase

In the analysed structure of *Mu* transposase, all three carboxylates of the conserved DDE triad were visible. The conserved glutamate was on a loop with its side chain pointing out into solvent. It was proposed that a simple rearrangement could bring the glutamate into the aspartate-containing active site region. The alignment confirms the identities of Asp-269 and Glu-392 and positively identifies Asp-336 as the three members of the *Mu* DDE motif (Rice & Mizuuchi, 1995) (Fig.1.2-2a).

The structure of ASV integrase

The ASV integrase structure provided an example of a metal ion bound at the active site (Bujacz et al., 1996). The segment containing the invariant glutamate (disordered in HIV integrase and forming loop inappropriately positioned for catalysis in *Mu* transposase) and the other two aspartate was well ordered, forming the active site (Bujacz et al., 1995, 1996) (Fig. 1.2-2c). This is the best evidence to show the three conserved carboxylate residues contribute to the active site by coordinating with the essential divalent ion.

1.3 The initial step of transposition - assembly of the active DNA-protein complex

As an initial step of transpositon, the transposase has to recognise and bind sequence specifically to the ends of the element, to assemble a synaptic DNA-protein complex. In this complex, the transposase binds to the transposase binding sites at the ends of the element by its DNA binding domain. Some host proteins and other factors might also be involved in this complex. The DNA-protein complexes are different from element to element, depending on how many transposase binding sites, how many transposase monomers, and how many host proteins and other factors are involved. However, only in this stable complex can the transposase be activated to catalyse the transposition reactions.

1.3.1 The transposase binding sites of different elements

The transposase binding sites, which are the DNA sequences to which the transposase binds, are different from element to element. These differences are not only reflected on their sequences, but also on their positions and numbers, probably reflecting the specific DNA-protein association required for the transposition of each element. The transposable elements can therefore be divided into two classes according to the positions of the binding sites: one with binding sites located within the inverted repeat, and the other with the binding sites located internally to the inverted repeat. In each class, the elements can be further divided into two groups depending on the numbers of the binding sites: one has only one binding site, and the other has more than one binding sites (Tab.1.3-1).

1.3.1.1 Elements with transposase binding sites located within the inverted repeats

For the transposase to recognise, then cut and fuse the transposon to a new site, it is a common idea that the transposase will first bind to the sequence near the ends of the element where the transposase carries out the cleavage and joining of the DNA strands. The terminal inverted repeats, which are asymmetrically located at both ends of the element, are very good places for the transposase to bind to. This has been shown to be the case for elements in this class. Most of them have a single binding site located towards the interior of the terminal inverted repeats. This suggests the existence of two functional domains in the terminal inverted repeats of these elements : a domain which is recognised and bound by the transposase, and the terminal two to three base pairs which are presumably positioned in the active site of the protein (Polard & Chandler, 1995)

Table 1.3-1 classification of transposable elements according to their transposase binding sites

	One binding site	More than one binding site
Binding site within IR	IS1, IS10, IS30 & IS 903 Tn3 Tc1 Mariner	Tc3
Binding site internally to IR	P element	Mu Tn7 En/Spm Ac

Elements with one binding site located within IR

The elements in this group have only one transposase binding site which is located within the terminal inverted repeat sequence of the elements. This class includes many bacterial insertion sequences, such as *IS1*, *IS10*, *IS50*, *IS30* and *IS903* (Derbyshire & Grindley, 1992; Vos et al., 1993; Wiegand & Reznikoff, 1994); *Tn3* (Ichikawa, et al., 1987 & 1990; Amemura, 1990), the *Tc1* from *C.elegans*; (Vos, et al., 1993; Colloms, et al., 1994) and *Mariner Himar1*from hore fly, *Haematobia irritans* (Lampe, et al., 1996).

For example, *Tc1* from *C. elegans* is a element of 1612 bp long, with 26 bp terminal inverted repeats and an open reading frame coding for the transposase Tc1A. Tc1A binds specifically to the inverted repeats . The Tc1A recognition site is located between base pair 5 and 26 from the end of *Tc1* (Vos, et al., 1993) (Fig.1.3-1a).

Elements with more than one binding sites located within IR

The transposases encoded by elements in this class also bind to the inverted repeat sequences of the elements, but bind to more than one position in each inverted repeat. *Tc3* element from *C. elegans* is an example.

Tc3 is 2335 bp long and has inverted repeats of 462 bp at each end. The *Tc3* transposase binds specifically to two regions within the 462 bp *Tc3* inverted repeat. One region is located at the end of the inverted repeat in between base pair 11 and 30, the other is located about 180 bp from the end in between base pair 184 and 202 (Colloms, et al., 1994)(Fig.1.3-1b).

1.3.1.2 Elements with transposase binding sites located internally to the inverted repeats

Although as it has been shown in 1.3.1.1 that the transposases of some elements interact sequence specifically with terminal inverted repeat sequences of the elements, the transposases from other elements tend to bind specifically to DNA sequences internally to the terminal inverted repeats. Most of those studied so far have more than one transposase binding site, indicating that they might form more complicated structure with the tranposase in the transposition process.

Elements with one binding site located internally to the IR

The *P* element of *Drosophila* is the only member of this group that has been identified so far. It has 31 bp inverted repeats and the *P* transposase binds to a 10 bp consensus sequence 16 bp from 5' inverted repeat and 4 bp from 3' inverted repeat (Kaufman et al., 1989)(Fig.1.3-1c).

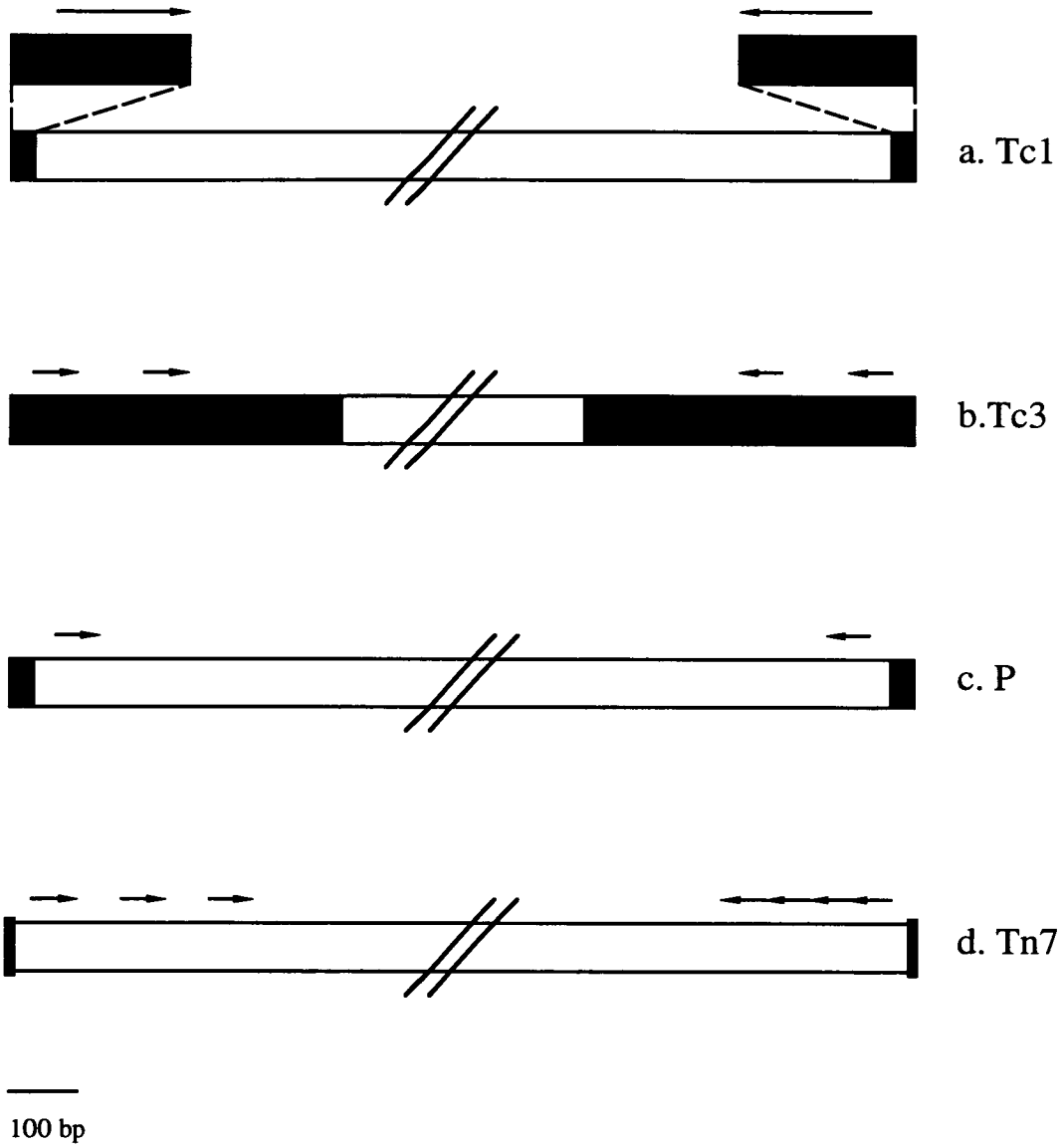


Fig. 1.3-1 The transposase binding sites of Tc1, Tc3, P, and Tn7 elements
The transposase binding sites are indicated by arrows.
The terminal inverted repeat regions are indicated in black blocks.

Elements with more than one binding sites located internally to the IR

This group includes most of the elements having binding sites located internally to the terminal inverted repeats. They have more than one binding sites at each end of the elements. These binding sites are normally highly conserved short sequences, about 10-20 bp in length. Their sequences and distributions are characteristic of each element, reflecting the specific DNA-protein interactions of different elements. This group includes *Mu*, *Tn7*, *En/Spm* and *Ac*.

For example, *Tn7* has 8 bp inverted repeats at both ends. These 8 bp inverted repeats are located outside of the outmost TnsB binding consensus sequences. In addition to these two outmost binding sites, five additional 22 bp consensus sequences for TnsB binding are located further inside, two on the left hand side, and three on the right hand side (Arciszewaka, et al., 1989)(Fig.1.3-1d).

1.3.2 The DNA binding domain and its functional motif of a transposase

1.3.2.1 General domains of a transposase

The transposase encoded by a transposable element is responsible for its transposition. The general duties of a transposase include recognition and binding to the transposon end sequence and the target DNA, to form an active DNA-protein complex, then catalysing a transposition reaction involving DNA cleavage and strand transfer. These duties have been shown to be shared by different domains of the transposase: a DNA binding domain, a catalytic domain, and a dimerisation domain.

DNA binding domain

The DNA binding domain of a transposase is normally located near the N-terminus of the transposase. It recognises and binds to the transposase binding site located near the ends of the element.

Catalytic domain

The catalytic domain of a transposase, which is normally located in the central region of the transposase, usually contains the conserved DDE catalytic motif, is responsible for catalysing the transposition reactions including DNA cleavage and strand transfer reactions.

Dimerisation domain

A stable synaptic DNA-protein complex is considered to be a prerequisite for the transposase to catalyse the transposition reaction. This is probably achieved by the association of two or more transposase monomers after they bind to the transposon end sequences. However, only a couple of transposases have been reported as having dimerisation domains. One of them is the transposase of *En/Spm* element of *Zea mays*, which has a dimerisation domain at the C-terminal of the transposase (Trentamann, et al., 1993). The other one is the transposase of *P* element. *P* transposase has a putative leucine zipper dimerisation motif located next to the N-terminal 98 amino acid DNA binding domain, this motif might represent a dimerisation domain involved in the transposase-transposase or transposase-repressor interactions (Rio, 1990; Lee et al., 1996).

1.3.2.2 The functional motifs in the DNA binding domains of DNA binding proteins

DNA binding proteins can be classified according to the structures of their functional motifs. These classes are referred to as helix-turn-helix, leucine zipper, helix-loop-helix, zinc finger and so on (Branden & Tooze, 1991; Harrison, 1991; Pabo & Sauer, 1992).

Helix-turn-helix (HTH)

This is the best characterised class among DNA binding proteins (Pabo & Sauer, 1984; Harrison & Aggarwal, 1990). The HTH is generally defined as a 20-residue segment, with two α–helices that cross at an angle of about 120°C. There may be additional helical residues, at the N-terminus of the first helix or the C-terminus of the second, which extend beyond the central 20 residue 'core' (Harrison, 1991). The second helix in the HTH motif is the "recognition" helix which penetrates into the major groove of the DNA target, allowing protein side-chains to make extensive base-specific hydrogen bonds and contacts (Brennan, 1991, 1992).

leucine zipper

The DNA binding domains of the leucine zipper proteins generally contain 60-80 residues (Landxchulz, et al., 1988; Hope & Struhl, 1986), and contain two distinct subdomains: the leucine zipper region mediates dimerisation, while a basic region contacts the DNA. Leucine zipper sequences are characterised by a heptad repeat of leucines over a region of 30-40 residues. There also tends to be a conserved repeat of

hydrophobic residues (often Val or Ile) occurring three residues to the N-terminal side of the leucines (Landxchulz, et al., 1988).

Helix-loop-helix (HLH)

The HLH proteins have a basic region that contacts the DNA and a neighbouring region that mediates dimer formation (Murre, et al., 1989a; Murre, et al., 1989b; Voronova & Baltimore, 1990). This dimerisation region forms an α–helix, a loop, and a second α–helix (Murre, et al., 1989a). The sequence of the basic region has some similarities with that of the basic region of the leucine zipper proteins (Prendergast & Ziff, 1989).

Zinc finger

Zinc fingers are another one of the major structural motifs involved in protein-DNA interactions (Berg, 1986; Klug & Rhodes, 1987; Kaptein, 1991; Berg, 1990). Proteins in this family usually contain tandem repeats of 30-residue zinc finger motif, in each motif with one Zn ion co-ordinated by two cysteines and two histidines, or four cysteines, or one histidines and three cysteines. The loop between these residues forms the DNA-binding region which binds into the major groove of DNA (Miller, et al., 1985; Brown, et al., 1985).

Since transposases are also DNA binding proteins, and their functions in transposition include specific DNA binding and dimerisation, the functional motifs described above might be expected to be present in transposases. Indeed, some of these motifs have been found in transposase sequences (see later), and probably play an important role in the transpositon.

1.3.2.3 The DNA binding domain and its functional motif of a transposase

DNA binding domains of some transposases

For a transposase to execute its catalytic activity in the transposition process, it has to first recognise and bind to the transposon end sequences to form a stable synaptic complex. Sequence specific DNA binding is a general characteristic of all transposases, and has been extensively studied in recent years. All the transposases studied so far have been show to be able to bind specifically to the transposon end sequences. The amino acid sequence responsible for this binding, the DNA binding domain, is always located at the N-terminal of the transposase.

Table 1.3-2 lists the DNA binding domains of the transposases studied so far. These include transposases of *IS30*, *Ac*, *P*, *Tc1*, *Tc3* elements (Stalder, et al., 1990; Lee et al., 1996; Vos & Plasterk, 1994; Collom et al., 1994).

Functional motifs in the DNA binding domains of some transposases

As DNA binding proteins, transposases probably have the same functional motifs as other DNA binding proteins. Indeed, some DNA binding motifs have been found in the DNA binding domains of *IS30*, *Tc1*, *Tc3* and *P* transposases.

***IS30* transposase**

The DNA binding domain of *IS30* element is located in the N-terminal 1/3 region of its transposase (Stalder, et al., 1990). A helix-turn-helix DNA-binding motif has been predicted to be present in the putative DNA binding region of this transposase family (Pietrokovski, 1996, Stalder, et al., 1990).

***Tc1-Tc3* transposase**

The DNA binding domain of *Tc1* transposase, Tc1A, is located in the N-terminal 68 amino acids of the transposase (Vos & Plasterk, 1994). The DNA binding domain of *Tc3* transposase is located at the 65 amino acids at the N-terminus of the transposase (Colloms, et al., 1994). Computer search based on three methods strongly indicated that the N-terminal region of *Tc1* and *Tc3* tranposase contains a HTH motif (Pietrikovski &Henikoff, 1997).

***P* transposase**

The DNA binding domain of *P* transposase has been mapped to the N-terminal 98 amino acids, which contains the potential CCHC ($CX_3CX_9HX_3C$) metal binding motif (Lee, et al., 1996; Rio, 1990).

1.3.3 The architecture of the synaptic DNA-protein complex

A well organised synaptic structure contains two transposon ends bound by the transposase is supposed to be a prerequisite for activating the transposase to catalyse the transposition reaction. The complex of each element has its own requirements for the number of transposase monomer, the number of the transposase binding site, and the kind of host proteins and other factors involved. The DNA-protein complexes of *Tn10*, phage *Mu*, and *Tn7* described here are good examples.

Table 1.3-2 The DNA binding domains of some transposases

Transposases	DNA binding domain	Reference
IS 30 transposase	N-terminal 1/3 region	Stalder, et al., 1990
Ac transposase	N-termianl 200 amino acids	Feldmar & Kunze, 1991
P transposase	N-terminal 98 amino acids	Lee, et al., 1996
Tc1 transposase	N-terminal 68 amino acids	Vos & Plasterk, 1994
Tc3 transposase	N-terminal 65amino acids	Colloms et al., 1994

The active DNA-protein complex of *Tn10*

Tn 10 is a composite bacterial transposon made up of two copies of the insertion sequence *IS10* that flank a tetracycline-resistance determinant (Kleckner, 1989; Kleckner et al., 1995). *IS10* right encodes the 46 kd *IS10* transposase. This transposase interacts specifically with the 23 bp terminal inverted repeats of *IS10*, and promotes transposition. The outside end of the element also has a specific binding site for Integration Host Factor (IHF) immediately internal to the terminal inverted repeat sequences (Huisman et al., 1989).

Transposition of this element occurs through a non-replicative mechanism in which the transposon is first excised from donor DNA by a pair of double-strand cleavages, then integrated into a new site by a single strand transfer (Bender & Kleckner, 1986; Haniford et al., 1989; Benjamin & Kleckner, 1992). An early and necessary step in *Tn10* transposition involves the pairing of two transposon ends. These two ends and the transposase make up a stable synaptic complex (Haniford et al., 1991; Haniford & Kleckner, 1994; Sakai et al., 1995). All the chemical steps (excision and strand transfer) in *Tn10* transposition take place in this synaptic complex (Sakai et al., 1995). A single *IS10* transposase monomer carries out all three chemical steps in the transposition reaction (Benjamin & Kleckner, 1992; Chalmers & Kleckner, 1994). Bacterial protein IHF serves as a regulatory element (Signon & Kleckner, 1995). Formation of a precleavage synaptic complex is followed by cleavage steps (Sakai et al., 1995; Hanifod et al., 1991). Cleavage of the two strands at each end occurs in a sequential order: nicking of the transferred strand always precedes nicking of the nontransferred strand (Balland & Kleckner, 1995). Once double-strand cleavage has occurred at both ends of the element, the synaptic complex is capable of capturing a target DNA *via* a stable noncovalent association; strand transfer then occurs (Bolland & Kleckner, 1996; Kleckner et al., 1995). It has recently been shown that all three chemical steps utilise a single active site containing critical amino acid residues. The same monomer promotes both DNA cleavages and strand transfer steps. Within any synaptic complex the two monomers carry out the entire chemistry of the transposition reaction (Bolland & Kleckner, 1996).

The active DNA-protein complex of phage *Mu*

Phage *Mu* transposes *via* a replicative mechanism, in which only the 3' end cleavage occurs, the transposon is attached to the target site by its 3' end in the strand transfer reaction while still remaining attached to the donor DNA by its 5' uncleaved end. This strand transfer product is processed later to generate two copies of the transposon fused to the donor backbone and the target DNA.

The *Mu* transposase, MuA, promotes the two chemical steps of the DNA cleavage and joining reactions required for transposition. Prior to initiation of these chemical steps, two *Mu* ends are held together by a MuA tetramer to form a stable synaptic complex. There are six transposase binding sites at the ends of *Mu* element: L1, L2 and L3 on the left end of the genome, and R1, R2 and R3 on the right end (Allison & Chaconas, 1992; Mizuuchi et al., 1992). These binding sites are related by a 22 bp consensus sequence (Craigie et al., 1984). L1 and R1 are closest to the ends of the genome. About 1 kb from the left end is an internal activating sequence (IAS) required for efficient *Mu* transposition (Mizuuchi & Mizuuchi, 1989; Liung et al., 1989; Surette et al., 1989). Before the synaptic complex forms, monomers of MuA protein bind to all the MuA binding sites; while in the synaptic complex, they are tightly bound to only the three endmost sites, L1, R1 and R2. and assemble into a tetramer which maintains the two DNA ends in a stablely synapsed complex (Mizuuchi, et al., 1992). The assembly of this synaptic complex require the presence of MuA transposase, divalent mental ions, bacterial proteins HU and IHF, supercoiled donor DNA containing the IAS, and properly oritated left and right end sequences (Mizuuchi, et al., 1992). The DNA cleavage step is processed after synaptic complex formation. The second chemical step, the DNA strand transfer reaction, converts the synaptic complex into the strand transfer complex (Surette et al., 1987). The internal MuA-binding sequences (L2, L3, and R3), the IAS, and the host proteins, HU and IHF, are not required for strand transfer (Mizuuchi et al., 1991; Lavoie & Chacona, 1990). The second phage-encoded transposition protein, MuB, which requires ATP for its action, strongly influences the choice of the target site by stimulating intermolecular strand transfer (Maxwell et al., 1987; Adzuma & Mizuuchi, 1988; Baker et al., 1991). The arrangement of MuA at the *Mu* ends in the second complex is similar to that in the first complex (Mizuuchi et al., 1991), although the second one contains the target DNA and MuB (Lavoie & Chaconas, 1990).

The active DNA-protein complex of *Tn7*

The bacterial transposon *Tn7* (Barth et al., 1976; Craig, 1996c) moves through a cut and paste mechanism involving formation of double-strand breaks at both ends of *Tn7* and joining of the 3' ends to the target DNA (Bainton et al., 1991, 1993). *Tn7* transposition involves multiple *Tn7* encoded proteins: TnsA, TnsB, TnsC, TnsD and TnsE. Different subsets of Tns proteins promote insertion into different target sites: TnsABC+D promote *Tn7* insertion into a specific site in the *E.coli* chromosome, *attTn7*, and Tns ABC+E promote *Tn7* insertion into many non-*attTn7* sites (Rogers et al., 1986; Waddell and Craig, 1988; Kubo & Craig, 1990). Molecular analysis of *Tn7*

transposition into *attTn7* revealed that all components of this reaction (TnsA, TnsB, TnsC, TnsD, the donor transposon DNA and the target *attTn7* DNA) must be present for the initiation of recombination (Bainton et al., 1991, 1993; Sarnovsky, et al., 1996). This requirement suggests that a prerequisite for recombination is the formation of an elaborate nucleo-protein complex involving specific recognition of the substrate DNAs, and also involving synapsis of the transposon ends with each other and with the target DNA; multiple specific protein-DNA and protein-protein interactions are surely necessary to form this active complex. *Tn7* transposase is a hetromeric complex of TnsA and TnsB, each protein executing different DNA processing reactions which act on different strands at the transposon end. TnsA mediates DNA cleavage reactions at the 5' ends of *Tn7*, and TnsB specially recognises the ends of *Tn7* and mediates DNA breakage and joining reactions at the 3' ends of *Tn7*. TnsB is the protein which executes the most central and fundamental of these breakage and joining reactions. Thus the double-strand breaks that underlie *Tn7* excision result from a collaboration between two active sites, one in TnsA and one in TnsB; the same (or a closely related) active site in TnsB also mediates the subsequent joining of the 3' ends to the target (Sarnovsky, et al., 1996).

1.3.3.4 The DNA-protein complexes of *Tn10*, *Mu* and *Tn7* - similarities and differences

The obvious similarity is that synaptic DNA-protein complexes are formed for all three elements as an initial step of their transpositions. Each of these complexes contains two transposon ends and the transposase. This complex has been shown to be crucial for the following transposition reactions in all these elements.

However, each complex is characteristic of each element, they are different in several aspects: a) the components in the complex are different. *Tn10* complex contains two transposase monomers, a pair of transposon ends, and host IHF protein; *Mu* complex contains a transposase tetramer, three transposase binding sites, IAS sequence, host HU and IHF proteins; while for *Tn7* complex, all components of the transposition reaction, TnsA, TnsB, TnsC, TnsD, the donor transposon DNA and the target *attTn7* DNA must be present for this initiation step. b) the properties of the transposases used to catalyse each transposition step are different. *IS10* transposase monomer catalyses all three steps of transposition using a single active site; for *Mu*, two MuA monomers within the tetramer catalyse the DNA cleavage reaction, the other two catalyse strand transfer reaction, a structure transfer is also required for changing the active site; for *Tn7*, TnsB catalyses 3' end cleavage and joining, while TnsA catalyses 5' cleavage. c) the interactions of the transposase monomer(s) are different.

IS10 transposase monomer is sufficient to catalyse the transposition reaction, in *Mu* transposition, a transposase tetramer is required, *Tn7* transposition needs a heteromeric transposase formed from TnsA and TnsB.

Taken together, *Tn10/IS10* transposition probably has the simplest requirements for initiating a transposition reaction, a transposase monomer catalyses all three chemical reactions; *Mu* has more complicated requirement, the co-operation of four transposase monomers is needed; while the most complicated *Tn7* transposition needs the co-operation of proteins of different gene products. The different requirements might reflect different levels of regulation in these transposition processes.

Chapter 2

MATERIAL AND METHODS

2.1 Materials

2.1.1 Chemicals

The chemicals used in this study were either products for molecular biology or analytical grade chemicals. They were supplied from companies of Sigma, Fisons, BDH, Boehringer Mannheim.

2.1.2 Enzymes

Restriction enzymes and their buffers were supplied from companies of Boehringer Mannheim or New England Biolabs, Inc.; Klenow enzyme and its buffer were from New England Biolabs, Inc. or Promega; T4 DNA ligase and its buffer were from New England Biolabs, Inc.; Taq DNA polymerase and its buffer were from GIBCOBRL or Promega.

2.1.3 Isotopes

α-^{32}P-dCTP (3,000 Ci/mmol)

α-^{35}S-dATP (400 Ci/mmol)

The isotopes used in this study were supplied from Amersham International.

2.1.4 Bacterial media

Luria Broth (L-broth):

Difco Bacto tryptone, 10 g; Difco Bacto yeast extract, 5 g; NaCl, 5 g; per litre, adjusted to pH 7.2 and autoclaved.

Luria agar (L-agar):

Same as Luria Broth with 15 g per litre Difco agar.

Ampicillin to a final concentration of 100 μg/ml was added to media immediately prior to use when required.

2.1.5 Solutions

5x Binding buffer:

100 mM HEPES, pH7.6; 200 mM KCl; 10 mM $MgCl_2$; 0.5 mM EDTA; 5 mM DTT; 50% glycerol.

Boiling mix:

Stacking gel buffer, 1.0 ml; 25% SDS, 0.8 ml; ß-mercaptoethanol, 0.5 ml; 100% glycerol, 1.0 ml; Bromophenol blue, 0.05%.

Denaturation buffer:

0.5 M NaOH; 1.5 M NaCl.

Destain:

Methanol,10%; Acetic acid,10%; dH_2O, 80%.

FDE loading buffer:

EDTA, 100 mM, pH 8.0; Bromophenol blue, 0.1%; Ficoll, 20%.

Gel Buffer:

Tris, 18.15g; SDS, 0.4g. pH to 8.9 with HCl; make up to 100 ml with dH_2O.

Glutathione-agarose bead slurry:

Preswell glutathion-agarose beads (Sigma) in 10 vol 1x PBS for 1 hr. Wash twice with 1x PBS and store as 50% slurry at 4°C for up to 1 month.

Hybridisation solution:

0.5 M NaP (1 M Na_2HPO_4, 72 ml; 1 M NaH_2PO_4, 28 ml; make up 1 M NaP); 7% SDS; 1 mM EDTA, pH 8.0.

Neutralisation buffer:

0.5 M Tris·HCl, pH 7.5; 1.5 M NaCl.

10x PBS:

NaCl, 80 g; KCl, 20g; Na_2HPO_4, 20 g; make up to 1 L with dH_2O.

Reaction stop buffer: 1 % SDS; 200 mM NaCl; 20 mM EDTA, pH 8.0; 40 µg/ml tRNA.

Sequencing loading buffer (Stop solution):

95% Formamide; 20 mM EDTA; 0.05% Bromophenol Blue; 0.05% Xylene Cyanol FF.

Stacking gel buffer:

Tris, 5.1g; SDS, 0.4g; pH to 6.7 with HCl; make up to 100 ml with dH_2O.

Stain:

4.0 g CBBR, 1.0 L methanol, 200 ml acetic acid, 1.0 L dH_2O.

SOC buffer:

L-broth, 100 ml; 20% Glucose, 1.8 ml; 1M $MgSO_4$, 1 ml; 1 M $MgCl_2$, 1 ml.

STET:

Sucrose, 8%; Triton X-100, 0.5%; 50 mM EDTA (pH 8.0); 50 mM Tris-HCl (pH 8.0); autoclaved.

10x TAE:

Tris, 48.4g; Acetic acid, 11.4 ml; 0.5 M EDTA (pH 8.0), 20 ml; make up to 1 L with dH_2O.

10xTBE:

Tris, 108 g; Boric acid, 55g; EDTA, 9.5g; make up to 1 L with dH_2O.

TE:

Tris, 10 mM; EDTA, 50 mM; pH 8.0. Autoclaved.

10xTGS:

Tris, 30.0g; Glycine, 144.0 g; SDS, 10.0 g; make up to 1 L with dH_2O.

TfbI buffer:

3 mM potassium acetate; 100 mM RbCl; 50 mM $MnCl_2.4H_2O$; 10 mM $CaCl_2.2H_2O$; 15% glycerol; pH to 5.8 with 0.2 M Acetic acid; Sterilise by filtration.

TfbII buffer:

10 mM MOPS; 10 mM RbCl; 75 mM $CaCl_2.2H_2O$; 15 % glycerol; pH to 6.8 with KOH; Sterilise by filtration.

Thrombin cleavage buffer:

2.5 mM $CaCl_2$ in wash buffer.

Wash buffer:

50 mM Tris-HCl (pH 7.5) / 150 mM NaCl.

Wash solution:

40 mM NaP (1 M Na_2HPO_4, 72 ml; 1 M NaH_2PO_4, 28 ml; make up 1 M NaP); 1% SDS; 1 mM EDTA, pH 8.0; make up to 1 L with dH_2O.

2.1.6 *E.coli* strains

NAME	GENOTYPE	REFERENCE
NM522	F' *lacI^q Δ(lacZ)M15 proA$^+$B$^+$ /supE thi Δ(lac-proAB) Δ(hsdMS-mcrB)5*($r_K^-m_K^-$ McrBC$^-$)	Woodcock, et al. (1989); Gough and Murray (1983)
XL-1 BLUE	F' :: Tn*10 proA$^+$B$^+$ lacI^q Δ(lacZ)M15 /recA1 endA1 gyrA96*(Nal^r) *thi hsdR17*($r_K^-m_K^+$) *supE44 relA1 lac*	Bullock *et al.* (1987)

2.1.7 Plasmids

NAME	DESCRIPTION	REFERENCE
pGEX4T-2	For inducible, high-level intracellular expression of genes or gene fragments as fusions with glutathion S-transferase (GST).	Smith, D. B. and Johnson, K. S. (1988)
pogoR11XC	The *ClaI-XhoI* fragment from pogoR11 was subcloned into plasmid pBR322 between the *ClaI* and *SalI* sites.	Tudor, M. *et al* (1992)

2.1.8 Oligonuleotides

NAME	SEQUENCES (5' to 3')	COMMENTS
M0804 (3'ORF1+5'ORF2)(-)	**ACTGAACTTAAATCCAGC CTTTTTAAAACAATTCTG**	- strand PCR primer from base +1420 to +1510 of *pogo* sequence with a deletion from base +1438 to +1496.
M0805 (3'ORF1+5'ORF2)(+)	**CAGAATTGTTTTAAAAAG GCTGGATTTAAGTTCAGT**	+strand PCR primer from base +1420 to +1510 of *pogo* sequencewith a deletion from base +1438 to +1496.
Oligo1 (20/2/96)	CCG*CTCGAG***GGCTGTTAT CTGCGATGCAGC**	- strand PCR primer from base +111 to +131 of *pogo* sequence.
Oligo2 (20/2/96)	CGC*GGATCC***CAGTATAATT CGCTTAGCTGCATCGATA GTTAGCTGCATCGGC**	+ strand PCR primer from base +1 to +43 of *pogo* sequence.
Oligo1 (11/3/96)	CGC*GGATCC***ATCGATAGTT AGCTGCATCGGC**	+ strand PCR primer from base +22 to +43 of *pogo* sequence.

Oligo2 (11/3/96)	CCG*CTCGAG***CAGCTGATA TGATGCACTTAAG**	- strand PCR primer from base +144 to +160 of *pogo* sequence.
Oligo1 (10/4/96)	GC*CTCGAG***CCTTTAAGCA ATCCTGGC**	- strand PCR primer from base +798 to +815 of *pogo* sequence.
Oligo2 (10/4/96)	GC*CTCGAG***GTAGCAGTTT CGCCGTG**	- strand PCR primer from base +739 to +755 of *pogo* sequence.
Oligo3 (10/4/96)	CGC*GGATCC***CATTAAGAAT GAAATTATCAGTTCC**	+ strand PCR primer from base +1960 to +1984 of *pogo* sequence.
Oligo4 (10/4/96)	CCG*CTCGAG***TACTTTGCAC AATGCC**	- strand PCR primer from base +2079 to +2100 of *pogo* sequence.
Oligo 1 (3/5/96)	GCG*GGATCC***TGGCGCTGG CGCAAGCGCC**	+ strand PCR primer from base +697 to +715 of *pogo* sequence.
Oligo 2 (3/5/96)	GC*GGATCC***GGTTATAACC CAGAAGAC**	+ strand PCR primer from base +814 to +831 of *pogo* sequence.
S5046 (ORF1-5')	GC*GGATCC***GGTAAAACAA AGCGTGTCG**	+ strand PCR primer from base +343 to +361 of *pogo* sequence.
S5047 (ORF1-3')	GC*CTCGAG***CAATAATAATC CCATCTTACC**	- strand PCR primer from base +1437 to +1458 of *pogo* sequence.
S5048 (ORF2-5')	GC*GGATCC***GCTGGATTTA AGTTCAG**	+ strand PCR primer from base +1493 to +1509 of *pogo* sequence.
S5049 (ORF2-3')	GC*CTCGAG***CTAACTTAAAA ATAGTCAG**	- strand PCR primer from base +1881 to +1899 of *pogo* sequence.
T6393	GC*TATAGA***TACAGTATAAT TCGCTTAGCTGCCTCGAG TACTTTGCACAATGCC**	- strand PCR primer from base +2079 to +2121 of *pogo* sequence.
T6394	GC*GGATCC***TACAGTATAAT TCGCTTAGCTGCATCGAT AGTTAGCTGCATCGGC**	+ strand PCR primer from base +1 to +43 of *pogo* sequence.
T6395 (3' N75)	GC*CTCGAG***CTACTAAGTC GTGCGCTCC**	- strand PCR primer from base +544 to +562 of *pogo* sequence.

T6396 (3' 95 probe)	GC*CTCGAG***CGTACAAATTT** **TCTATTCACAC**	- strand PCR primer from base +74 to +95 of *pogo* sequence.
T7175 (5' 43 probe)	CCG*CTCGAG***GCCGATGCA** **GCTAACTATCGATGCAGC** **TAAGCGAATTATACTG**	- strand PCR primer from base +1 to +43 of *pogo* sequence.
T7880 (5' Δ43 probe)	GC*GGATCC***AAGATATCTG** **CATTATTTTTCC**	+ strand PCR primer from base +44 to +65 of *pogo* sequence.
V5497 (pogo R11-3')	TATATGTATATGTATGTATAT ACTA**CAG**	- strand PCR primer including 25 bp pBR322 sequence downstream of pogo sequence in pogo R11 XC and +2119 to +2121 of *pogo* sequence.
V5498 (pogo R11-5')	ATCGACCTGTTGGCCATATA GTATA**CAG**	+ strand PCR primer including 25 bp pBR322 sequence upstream of pogo sequence in pogo R11 XC and +1 to +3 of *pogo* sequence.
V5734 (5'-ΔN78)	GC*GGATCC***GAAGAAGCCT** **TATACATTTGG**	+ strand PCR primer from 562 to 582 of *pogo* sequence.
V6198 (Bam-R11-5')	GC*GGATCC*ATCGACCTGTTG GCCATATAG	+ strand PCR primer including 21 bp pBR322 sequence upstream of *pogo* sequence in pogo R11XC.
V6333 (3' of 5'IR)	GC*CTCGAG***GCAGCTAAGC** **GAATTATACTG**	- strand PCR primer from +1 to +21 of *pogo* sequence.
V7013 [R44 (+)]	**CCACAGTCAACGCCATTT** **TACAAAAAAC**	+ strand PCR primer from base +458 to +485 of *pogo* sequence for mutating *pogo* transposase R44 from Arginine to Alanine.
V7014 [R44 (-)]	**GTTTTTTGTAAAATGGCG** **TTGACTGTGG**	- strand PCR primer from base +458 to +485 of *pogo* sequence for mutating *pogo* transposase R44 from Arginine to Alanine.
V7015 [K48 (+)]	**GCATTTTACAAGCCACAA** **ATGAAATTC**	+ strand PCR primer from base +470 to +496 of *pogo* sequence for mutating *pogo* transposase K48 from Lysine to Alanine.

V7016 [K48 (-)]	**GAATTTCATTTGTGGCTT GTAAAATGC**	- strand PCR primer from base +470 to +496 of *pogo* sequence for mutating *pogo* transposase K48 from Lysine to Alanine.
V7017 [R39 (+)]	**CAAATGCGACGCCTCCA CAGTC**	+ strand PCR primer from base +444 to +465 of *pogo* sequence for mutating *pogo* transposase R39 from Arginine to Alanine.
V7018 [R39 (-)]	**GACTGTGGAGGCGTCGC ATTTG**	- strand PCR primer from base +444 to +465 of *pogo* sequence for mutating *pogo* transposase R39 from Arginine to Alanine.
V7459 [Δ(22-43)(+)]	**CGCTTAGCTGCAAGATAT CTGC**	+ strand oligo nucleotides from +11 to +21 plus +42 to +54 of *pogo* sequence.
V7460 [Δ(22-43)(-)]	**GCAGATATCTTGCAGCTA AGCG**	- strand oligo nucleotide from +11 to +21 plus +42 to +54 of *pogo* sequence.
V7461 [(22-43)(-)]	*CGTAGC***CGATGCAGCTAA CTATCGAT***GGATCC*	- strand oligo nuclotide from 22 to 43 of *pogo* sequence.
V7912 [C32-P(+)]	**CAAAAAGGAAATTCCCG CCAAGTTCAAATG**	+ strand PCR primer from base +420 to +449 of *pogo* sequence for mutating *pogo* transposase C32 from Cysteine to Proline.
V7913 [C32-P(-)]	**CATTTGAACTTGGCGGG AATTTCCTTTTTG**	- strand PCR primer from base +420 to +449 of *pogo* sequence for mutating *pogo* transposase C32 from Cysteine to Proline.
V7914 [V42-P(+)]	**GACAGATCCACACCCAA CCGCATTTTAC**	+ strand PCR primer from base +451 to +478 of *pogo* sequence for mutating *pogo* transposase V42 from Valine to Proline.

V7915 [V42-P(-)]	**GTAAAATGCGGTTGGGT GTGGATCTGTC**	- strand PCR primer from base +451 to +478 of *pogo* sequence for mutating *pogo* transposase V42 from Valine to Proline.
V8599 [3'BS12(+)]	GC*GGATCC***GGCATTGTGC AAAGTACTCG**	+ strand PCR primer from base +2079 to +2098 of *pogo* sequence.
V8600 [3'BS 12 (-)]	GC*GGATCC***CAGTATAATTC GCTTAGCTGC**	- strand PCR primer from base +2101 to 2121 of *pogo* sequence.
V8601 [BS 14 (+)]	GATCCGAATTCGTCGAC**AG TTAGCTGCATCG**TCTAGAC TGCAGC	+ strand oligo nucleotide containing the 14 bp pogo transposase binding site from base +28 to +41 of *pogo* sequence.
V9045 [BS 14 (-)]	TCGAGCTGCAGTCTAGA**CG ATGCAGCTAACT**GTCGAC GAATTCG	- strand oligo nucleotide containing the 14 bp pogo transposase binding site from base +28 to +41 of *pogo* sequence.
W0914 [K34A (+)]	**GGAAATTTGTGCCGCCTT CAAATGCGACAG**	+ strand PCR primer from base +426 to +455 of *pogo* sequence for mutating *pogo* transposase R34 from Lysine to Alanine.
W0915 [K34A (-)]	**CTGTCGCATTTGAAGGC GGCACAAATTTCC**	- strand PCR primer from base +426 to +455 of *pogo* sequence for mutating *pogo* transposase R34 from Lysine to Alanine.
W0916 [R63A (+)]	**CGTCAGGTTTAAAAGCCA AGCGTCAAAGAAAAGG**	+ strand PCR primer from base +526 to +528 of *pogo* sequence for mutating *pogo* transposase R63 from Arginine to Alanine.
W0917 [R63A (-)]	**CCTTTTCTTTGACGCTTG GCTTTTAAACCTGACG**	- strand PCR primer from base +526 to +528 of *pogo* sequence for mutating *pogo* transposase R63 from Arginine to Alanine.

pogo sequences are shown in bold letters.

Restriction site are shown in italic.

Extra bases are show in plain.
Mutated bases are underlined.

2.2 Methods

2.2.1 Manipulation of bacteria

2.2.1.1 Growth of bacterial cultures

Liquid culture was used to multiply large amounts of bacteria. A single colony of *E.coli* was inoculated using a sterile inoculation loop to L-broth, supplemented with 100 μg/ml of ampicillin or other antibiotics when required, and was grown at 37°C with vigorous shaking for overnight. Cultures were normally grown in conical flasks or universal bottles with a capacity of 5 times of the culture.

2.2.1.2 Storage of bacterial cultures

A single colony of *E.coli* was picked up with a sterile needle and stabbed through L-agar in a small glass vial. After overnight growth at 37°C, the stab was stored at room temperature. Alternatively, the culture can be stored for longer by mixing up fresh overnight bacterial culture with equal volume of glycerol, normally 1 ml each, and stored in a sterile vial at -70°C. For recovery, the culture was streaked out from either stabs or thawed frozen culture onto a L-agar plate, with antibiotic when required.

2.2.1.3 Preparation of competent *E.coli* cells

Calcium chloride method

This method was modified from Mandel and Higa, 1970.

An overnight bacterial culture was diluted 1 in 20 and grown with vigorous shaking at 37°C to a density of ~$5x10^7$ cells/ml, which corresponds to a OD_{550} of 0.2 for a *rec+* strain and OD_{550} of 0.5 for a *rec-* strain. The culture was chilled on ice for 5 min and centrifuged at 4000 rpm for 10 min at 4 °C to pellet the cell. The supernatant was discarded, the cells were resuspended in half of the original culture volume of an ice-cold, sterile solution of 50 mM $CaCl_2$ and 10 mM Tris·HCl (pH 8.0). The cell suspension was placed in ice for 15 min and then centrifuged at 4000 rpm for 10 min at 4 °C. The supernatant was discarded, the cells were resuspended in 1/15 of the original volume of an ice-cold, sterile solution of 50 mM $CaCl_2$ and 10

mM Tris·HCl (pH 8.0). The cell suspension was dispensed as 0.2 ml aliquots into prechilled eppendorfs and stored at 4°C for 12-24 hr.

For maximum transformation efficiency, it is very important (1) that the bacterial culture is in the logarithmic phase and that the cell density is low at the time of treatment with calcium chloride; and (2) that the cells are maintained at 4°C for 12-24 hr. During this period, the efficiency of transformation increases 4 to 6 folds (Dagret and Ehrlich 1979). After an additional 24 hr, the efficiency decreases to the original level.

Rubidium chloride method

This method was modified from Hanahan, D., 1985.

An overnight bacterial culture was diluted 1 in 20 and grown with vigorous shaking at 37°C to a density of 4-7x10^7 cells/ml, which corresponds to a culture of *rec-* strains at OD_{550} of 0.35-0.60. The culture was chilled on ice for 15 min then the cells were pelleted by centrifugation at 6000 rpm for 5 min at 4 °C. The supernatant was removed thoroughly, the cell pellet was resuspended in 1/3 of the original culture volume of TfbI buffer and incubated on ice for 15 min. The cells were pelleted by centrifugation at 6000 rpm for 5 min, resuspended in 1/12.5 of the original culture volume of TfbII buffer and incubated on ice for 15 min. The cell suspension was aliquoted as 0.1 ml each into 1.5 ml eppendorfs and stored at -70 °C.

2.2.1.4 Transformation

Heatshock method

1-10 µl of plasmid DNA in ligation buffer or TE buffer was added to 30-100 µl of competent cells thawed slowly on ice and incubated on ice for 30 min. The mixture was incubated at 42 °C for 2 min and then back onto ice for 2 min. 80 µl of L-broth was added into the mixture and incubated at 37 °C for 30-60 min. The culture was spread onto a suitable selective plate and incubate at 37 °C overnight.

For blue/white selection, 100 µl 100 mM IPTG and 20 µl 50 mg/ml X-gel were added to the culture before it was spread to the plate.

Electropration method

This method was modified from Heary & Duncan, 1989. Transformation frequency was increased by using it.

A single colony of *E.coli* was inoculated onto 10 ml of 2xYT media and grown at 37°C overnight with vigorous shaking. The cells were chilled on ice, harvested by

centrifugation at 4,000 rpm for 10 min at 4 °C. Then the cells were washed 3 times by resuspending the pellet in 10 ml of water followed by centrifugation. Finally the cells were resuspended in approximate 150 µl of water.

2 µl of ligated DNA was mixed with 40 µl of cells in an eppendorf then transferred to the bottom of a electroporation cuvette. The BIO-RAD GENE Pulser™ electroporator machine was set to 25 µF, 2.5 KV and 200 Ω. A cuvette was placed in the slide and pushed in. The two red buttons were pressed until 'ping' signals completed. The cuvette was removed quickly and 1 ml of SOC solution, which had been sucked in a Gilson tip, was added into the cuvette immediately and mixed with the cells. The mixture was transferred to a small glass bottle and grown at 37 °C for 20-60 min before plated out on a suitable selecting plate.

2.2.2 Nucleic acid preparation and manipulation techniques

2.2.2.1 Small scale preparation of plasmid DNA

Boiling miniprep

This method is modified from the boiling method of Holmes and Quigley, 1981.

A single plasmid carrying colony was grown overnight at 37 °C in 5 ml of L-broth containing suitable selective antibiotic. Cells from 1.5 ml of overnight culture were pelleted by centrifugation in a table-top microcentrifuge for 2-5 min. The pellet was resuspended in 200 µl of STET and 20 µl of lysozyme (10 mg/ml). The suspension was boiled at 100 °C for 40 sec then centrifuged for 10 min. After the soft pellet was removed, 200 µl of isopropanol and 20 µl of 3M sodium acetate were added to the supernatant. The suspension was incubated at -70 °C for 15 min then centrifuged for 10 min to precipitate the plasmid DNA. The DNA pellet was washed once with ether, dried at room temperature before resuspended in 50 µl of TE.

QIAGEN miniprep

Small scale of plasmid DNA was prepared from 1-5 ml of *E.coli* overnight culture by using QIAGEN miniprep kit supplied by QIAGEN. Supplier's instructions were followed.

2.2.2.2 Large scale preparation of plasmid DNA (QIAGEN Midiprep)

Large scale of plasmid DNA was prepared from 25-100 ml of *E.coli* overnight culture by using QIAGEN midiprep kit supplied by QIAGEN. Supplier's instructions were followed.

2.2.2.3 Quantification of nucleic acid concentration

The concentration of a DNA sample could be roughly estimated by running the sample at the same time with a standard DNA on an agarose gel and comparing the intensity of the bends of the two DNA samples.

More accurately, the concentration of a DNA sample could be measured by spectophotometry. Either the Lambda 15 UV/VIS Spectrophotometer (PERKIN ELMER) or the Gene Quant (Pharmacia Biotech) was used. Approximately 50 μg/ml of double-stranded DNA gives 1 OD_{260}.

2.2.2.4 Phenol extraction of proteins from DNA

Distilled phenol was equilibrated to pH 8.0 by repeatedly mixing with an equal volume of 1 M Tris·HCl (pH8.0). DNA to be extracted was mixed thoroughly with 1 volume of phenol then centrifuged for 5 min. The upper aqueous phase containing the DNA was removed to a fresh tube. Instead of , or following phenol extraction, DNA was extracted with phenol:chloroform:isoamylalcohol (25:24:1). The upper aqueous phase was then extracted with an equal volume of chloroform:isoamylalcohol (24:1) to remove any remaining phenol. After spinning the upper aqueous phase was transferred to a fresh tube.

2.2.2.5 Ethanol precipitation of DNA

DNA was precipitated from solution by addition of 1/10 volume of 3 M Sodium Acetate (pH 5.8) and 2 volume of ethanol. The solution was mixed thoroughly and incubated at -20 °C overnight or -70 °C for 15 min. DNA was pelleted by centrifugation at high speed for 15 min. After discarding the supernatant, the pellet was washed with 70% ethanol and spun again for 2 min. The supernatant was removed and the DNA pellet was dried under vacuum or at room temperature before dissolved in TE.

2.2.2.6 Agarose gel electrophoresis

Agarose gel electrophoresis was used for identification or separation of DNA or DNA fragments following purification or restriction enzyme digestion. DNA samples were mixed with 1/6 volume of FDE loading buffer before loading onto 0.7-

2.0% agarose gel, depending the size of DNA. The gel can be prepared and run in either 1x TAE or 1x TBE buffer. Ethidium bromide stock solution (10 mg/ml) was added to the gel to make a final concentration of 1 μg/ml. Electrophoresis was performed horizontally at a difference of approximately 10 V/cm. The bands of DNA could be visualised under UV illumination and photographed.

2.2.2.7 Cleavage of DNA with restriction enzymes

DNA was digested by restriction enzyme in 20-100 μl of appropriate restriction buffer. At least 1 unit of enzyme was used for digesting 1 μg of DNA at 37°C for 1 hour. To ensure complete digestion, excess amount of enzyme or longer incubation period was normally used.

2.2.2.8 Recovery of DNA from agarose gels

GENECLEAN

DNA fragments larger than 200 bp were purified from agarose gel or solution by using GENECLEAN II kit supplied by BIO 101 Inc. Supplier's instructions were followed.

MERmaid

DNA fragments between 10 bp and 200 bp were purified from agarose gel or solution by using MERmaid kit supplied by BIO 101 Inc. Supplier's instructions were followed.

2.2.2.9 Ligation of DNA fragments to vector DNA

Vector DNA and the DNA fragment to be cloned were digested with the appropriate restriction enzyme and purified by ethanol precipitation or GENECLEAN/MERmaid. The vector and the DNA fragment of 4-8 times molar excess were ligated in 1x ligation buffer (10 mM Tris·HCl pH 7.2, 1mM EDTA, 10 mM $MgCl_2$, 10 mM DTT, 1 mM ATP) in the presence of 1 unit of T4 DNA ligase. This reaction was normally proceeded in a volume of 10 μl at room temperature for 3 hours or at 16°C for overnight.

2.2.2.10 Polymerase chain reaction (PCR)

PCR was used to amplify a specific DNA fragment from a template DNA.

A pair of primers were designed according to the sequences at the ends of the fragment to be amplified. The primers normally have 17-20 bases identical to the

template sequence depending on the G+C content of that region. A restriction site can be designed into 5' end of the primer for subsequent cloning. 2-3 extra G or C bases were normally designed at 5' end of the restriction site to facilitate restriction digestion. The 3' end of the primer was normally ended with G or C so that the primer could annealed better with the template, therefore facilitate extension of DNA synthesis.

The PCR reaction was set up in a 0.5 ml Eppendorf by mixing 100 nanogram of template DNA with 10 µl of 10x PCR buffer [200 mM Tris-HCl (pH 8.4), 500 mM KCl], 0.2 mM of each dNTP, 1 µM of each primer, 1.5 mM of $MgCl_2$, 0.05% (v/v) of W-1, 5 units of *Taq* polymerase (GIBCO BRL) and distilled water to make up to 100 µl. 75 µl of mineral oil was added to the top to prevent evaporation during the reaction.

The PCR program normally started by incubating the tube at 94 °C for 3 min to completely denature the template, followed by 25-35 cycles of amplification: Denaturation at 94 °C for 45 sec; Annealing at 55 °C for 30 sec; Extension at 72 °C for 1.5 min. Finally the tube was incubated at 72 °C for additional 10 min before the amplified product being analysed by agarose gel electrophoresis. The program can be varied with different reaction depends on the homology of the primer with the template and the length of the fragment to be amplified.

2.2.2.11 DNA sequencing by dideoxynucleotide method

Double stranded plasmid DNA was sequenced by using Sequenase™ version 2.0 kit supplied by USB (United States Biochemicals). Supplier's instructions were followed with some modification.

3-5 µg of miniprep DNA was denatured in 0.2 M EDTA, 0.2 M NaOH for 30 min at 37°C. Standard annealing and termination reaction were modified by addition of DMSO to a final concentration of 10% and 7.5% respectively. Annealing reactions were performed at 37°C for 30 min before chilling on ice. Labelling reactions were performed at room temperature for 5 min and termination reaction at 37 °C for 10 min.

Samples were loaded on to a 6% polyacrylamide denaturing gel after the gel being pre-run at 42 W to a temperature of approximate 50°C. The gel was run at 38-42 W to keep the best temperature until the dye reached a proper place. The gel was fixed in a solution containing 10% methanol and 10% acetic acid for 15 min, dried under vacuum at 80 °C and autoradiographed at room temperature for overnight.

2.2.2.12 Southern blotting

Southern blotting was used to transfer DNA bands from a gel to a nitrocellulose membrane after electrophoresis.

Two sheets of blotting paper were placed on a support with both ends in a tray of 20x SSC. The gel was placed on top of the paper. A nitrocellulose membrane (Hybond N, Amersham), the same size of the gel, was placed on top of it, followed by two sheets of 2x SSC soaked blotting paper, two sheets of dry blotting paper, a stack of paper towels and a weight. The transfer was allowed to proceed overnight before the membrane was rinsed in 2x SSC then UV cross linked to bind the DNA to the membrane.

2.2.2.13 Radioactive random primed labelling of DNA

DNA fragments were radioactive random primed labelled to be used as probes in hybridisation experiments. The DNA fragments were labelled with α-^{32}P-dCTP by using T7 Quickprime™ kit supplied by Pharmacia Biotech. Supplier's instructions were followed.

The labelled DNA was separated from unincorporated DNA by passing through the Nuctrap™ Push Column (STRATAGENE). Supplier's instructions were followed with some modification. The column was first prewetted with 70 µl of TE, then the DNA sample, which was made up to 70 µl with TE ,was applied. The labelled DNA was eluted from the column by 1-2 times wash with 70 µl TE.

The labelled DNA could also be separated from unincorporated DNA by using QIAquick PCR purification kit supplied by QIAGEN. Supplier's instructions were followed.

2.2.2.14 Radioactive end labelling of DNA

DNA fragments were radioactive end labelled to be used as probes in gel retardation experiments.

The DNA fragment to be labelled was digested with a restriction enzyme to create a cohesive end at one end. The 3' end of one of the strand was labelled when the cohesive end was filled in with α-^{32}P-dCTP and other dNTPs catalysed by Klenow polymerase.

200 ng of purified DNA was mixed with 3 µl of 10x labelling buffer, 3 µl of 20 mM dGTP, dATP and dTTP, 2 µl of α-^{32}P-dCTP (20 µCi) and 5 units of Klenow enzyme, made up to 30 µl with dH_2O. The mixture was incubated at room temperature for 30 min before 5 µl of ice-cold 5 mM dCTP was added. The reaction was allowed to continue for another 30 min at the same temperature.

The labelled DNA was separated from unincorporated DNA by using the methods as described in 2.2.2.13.

2.2.2.15 Hybridisation

Nitrocellulose membranes, with bound DNA, were placed in a Techne hybridisation cylinder and 10-20 ml of prewarmed hybridisation solution was added. The membranes were prehybridised for at least 1 hr at the hybridisation temperature, usually 65 °C for Southerns, in Techne hybridisation ovens.

The probe was denatured before being added to the cylinder. The hybridisation was proceeded for overnight. The membranes were washed three times with wash solution for 30 min each at a temperature 5 °C lower than hybridisation, before being wrapped in Saran wrap and placed in autoradiographic cassettes.

The radioactive signal was detected on X-ray films laid on top of the membrane. The films were stored at -70 °C for an appropriate length of time before being developed in an automatic X-ray film processor.

Probes could be kept in hybridisation solution at room temperature and then reused, after boiling for 20 min, to replace the prehybridisation solution.

Membranes could be stripped of probe DNA by boiling for 3 min in 0.1% SDS, then be reprobed.

2.2.2.16 Colony hybridisation

To identify bacteria harbouring recombinant plasmids, colonies were transferred to nitrocellulose discs by placing the filters onto the plates for 2 min. The discs were soaked in denaturation buffer for 2 min to lyse the cells and denature the DNA, neutralised in neutralisation buffer for 2 min, washed once with 2x SSC plus 0.1% SDS for 2 min, then twice with 2x SSC for 2 min each. The discs were UV cross linked before hybridisation was proceeded.

2.2.2.17 Autoradiography

Autoradiography was carried out by exposing membrane containing radioactive samples to a X-ray film in an autoradiographic cassette for overnight or longer. The exposure was carried out at -70 °C for α-^{32}P labelled samples such as southern blot membranes, and at room temperature for α-^{35}S labelled samples such as sequencing gels.

2.2.3 Protein preparation and manipulation techniques

2.2.3.1 Expression and purification of Glutathione-S-Transfrerase fusion proteins

The chosen DNA fragment was subcloned into the appropriate pGEX vector in the correct reading frame, competent *E coli* cells were transformed, and transformants were selected on L-Amp plates after overnight incubation at 37 °C .

Transformant colonies were picked into 10 ml L-Amp and grown at 37 °C with vigorous agitation until OD_{550} reached 0.3-0.5. The fusion protein expression was induced by adding 1 M IPTG to a final concentration of 0.1- 1.0 mM and continued incubation for another 2-5 hr.

1.5 ml of liquid culture was transferred to an eppendorf and centrifuged to pellet the cell. The pellet was resuspended in 300 µl of ice-cold PBS. Optionally, Triton X-100 was added to a final of 1% to minimise association of fusion protein with bacterial protein at the next step. The cells were lysed by using a probe sonicator for 10-30 sec. The insoluble material was removed by centrifugation for 5 min at 4°C and the supernatant was transferred to a fresh tube.

50 µl of 50% slurry of glutathione-agarose beads was added to the supernatant and the binding was carried on at 4 °C for 0.5-2 hr. The beads were collected by centrifugation for a few seconds and the supernatant was discarded. The beads were washed 3 times with 1 ml of PBS.

To identify the protein, 10 µl of boiling mix was added to the beads and the mixture was boiled at 100 °C for 3 min. The supernatant was loaded onto a 10-15 % SDS-polyacrylamide gel and run at 200 v for the appropriate time. The gel was then stained with stain solution and destained by destain solution to visualise the fusion protein.

To harvest the protein, the fusion protein was eluted by using 1 bead volume of freshly made 50 mM Tris.HCl (pH 8.0) containing 5 mM reduced glutathione (Sigma) (final pH 7.5) to wash the beads twice and collected the supernatant. The eluted protein was also analysed by SDS PAGE.

2.2.3.2 Cleavage GST tag from GST fusion protein

The glutathione-agarose beads bearing the fusion protein were washed twice with 20 vol of 1% Triton X-100 in 1x PBS. The beads were equilibrated by being washed once with wash buffer, then once with thrombin cleavage buffer. Finally the beads were resuspended in a small volume of thrombin cleavage buffer. Proper

amount of thrombin (0.2-1% wt/wt fusion protein) was added to the slurry and the mixture was incubated for 1 hr at room temperature. The released protein was recovered by washing beads with 1 bed vol of wash buffer and analysed by SDS-PAGE.

2.2.3.3 Quantification of protein concentration

The concentration of a protein sample could be determined by using Coomassie Protein Assay Reagent from PIERCE.

This reagent is based on the absorbence shift from 465 to 595 nm that occurs when Coomassie Brilliant Blue G-250 binds to proteins in an acidic solution. Since the colour response is non-linear over a wide range of protein concentration, a standard curve has to be prepared with each assay.

A BSA standard series which covered the range of concentration between 1 and 25 μg/ml was prepared, and each of them was mixed with Coomassie Protein Assay Reagent followed by reading the absorbance at 595 nm within 90 min. All assays were performed in duplicate. A standard curve was made by plotting the average net absorbance at 595 nm for each dilute BSA. The absorbance of an unknown protein sample was determined in the same way, and the concentration was determined by using the standard BSA curve. Supplier's instructions was followed.

2.2.3.4 Gel retardation assay

0.5-5 μg of protein was incubated with 1 μg of non-specific competitor poly dI-dC on ice in binding buffer for 10 min. 2 ng of DNA probe was added to the mixture and the incubation continued for another 20 min. In competition assay, 10-100 fold molar excess of unlabelled competitor DNA was added to the reaction at the same time as poly dI-dC. 1-2 μl of loading buffer was added to each sample before they were loaded to a 5% polyacrylamde gel, run at 150 v until the blue dye reached the bottom of the gel. The gel was then dried under vacuum and autoradiographed.

2.2.3.5 Excision assay

This method was from Vos, et al., 1996, with some modification.

Excision assay was used to determine the endonuclease activity of a transposase. 100-200 ng of the supercoiled plasmid DNA containing the transposon with the terminal inverted repeats was incubated with 50-100 ng of the transposase in a buffer containing 25 mM Tris (pH8.0), 25 mM NaCl, 1 mM DTT, 10% ethylene glycol, 5 mM $MgCl_2$ (or $MnCl_2$), 4 mM spermidine, 0.05% μg/μl of BSA for 15, 30, 45, and 60 minutes at 30 °C or 37 °C. After the reactions, the samples were run in an agarose

gel and visualised by ethidium bromide staining and southern blotting using the transposon DNA as probe.

RESULTS

Chapter 3

OVEREXPRESSION OF *POGO* TRANSPOSASE IN *E. COLI*

3.1 Introduction

The complete *pogo* element is 2121 bp long, it has two open reading frames of 1155bp and 417 bp on one strand, which, as indicated by cDNA analysis, are joined by RNA splicing to encode a putative *pogo* transposase (Tudor et al. 1992).

To test the above prediction, and, if possible, produce relatively large amounts of *pogo* transposase for activity studies, the DNA coding for the *pogo* transposase was amplified by PCR, and cloned into the pGEX4T-2 vector. This should express *pogo* transposase in *E. coli* fused to the C-terminus of glutathione S-transferase. The stability, the solubility of the expressed protein and the factors that affects its solubility were investigated. The approach to release the free transposase by protease cleavage of GST tag from the fusion protein was also studied in this chapter.

3.2 Results

3.2.1 Making pGEX-*pogo* construct

3.2.1.1 Design primers

Four primers (Fig.3.1; primer 1, S5046; primer 2, M0804; primer 3, M0805; primer 4, S5049; as described in 2.1.8) were designed to amplify and join the two open reading frames coding for *pogo* transposase by PCR. Primer 1 and 2 were used to amplify ORF1 while primer 3 and 4 were used to amplify ORF2. Primer 2 and 3 were designed according to the sequences at the 3' end of ORF1 and 5' end of ORF2, so that the intron in between these two ORFs could be deleted by PCR (Fig. 3.1). A *BamHI* and a *XhoI* restriction site were designed at the 5' ends of primer 1 and primer 4, respectively, for subsequent cloning. Two extra bases were added to the 5' of the restriction sites to facilitate restriction digestion.

3.2.1.2 Amplification and joining *pogo* ORF1 and ORF2 by PCR

ORF1 and ORF2 of *pogo* element were first amplified separately from pogo R11XC by PCR. pogoR11XC is a recombinant pBR322 plasmid containing the *ClaI-XhoI* fragment from pogoR11, between *ClaI* and *SalI* sites. The *ClaI-XhoI* fragment from pogoR11 contains a complete *pogo* element except for the 21 bp terminal inverted repeats at both ends (Tudor, et al, 1992). Two independent PCRs were conducted by using primer 1 and 2 for amplifying ORF 1, primer 3 and 4 for ORF2. These two PCR products were checked on an agarose gel before being mixed with equal molar and used as templates for the second PCR. The second PCR was conducted by using primer 1 and 4 and a mixture of ORF1 and ORF2 from previous PCR as templates

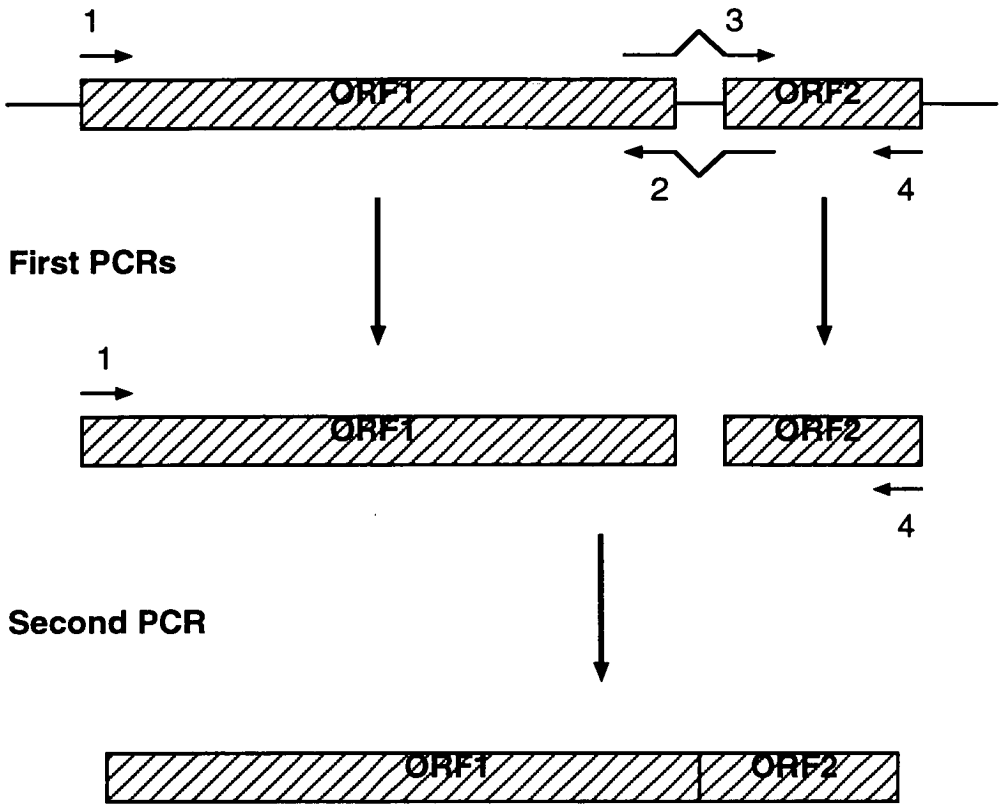


Fig. 3.1 Amplification and joining of pogo ORF1 and ORF2 by PCR
Four primers were designed to amplify and join the two ORFs from pogoR11XC by PCR. Primer1 and 2 were used to amplify ORF1 while primer 3 and 4 for ORF2. Primer2 and 3 were designed according to the sequences of 3' end of ORF1 and 5' end of ORF2, so that the intron in between the two ORFs could be deleted by PCR. Firstly, two indepedent PCR reactions were conducted to amplify ORF1 and ORF2 seperately. A second PCR was conducted by using equimolar amounts of the two first PCR products as the template and primer1 and 4 as the primers to obtain the joined ORF1-ORF2 product.

(Fig. 3.1). The final PCR product was purified by GENECLEAN and checked on an agarose gel before being used for subsequent cloning.

3.2.1.3 Cloning *pogo* ORF1-ORF2 into a pGEX vector

pGEX4T-2 vector was chosen for inducible, high-level intracellular expression of *pogo* ORF1-ORF2 as a fusion protein to the C-terminus of glutathione S-transferase (GST), a 27.6 kD cytoplasmic protein of eukaryotes. The fusion protein can be easily purified from bacterial lysates under nondenaturing conditions by affinity chromatography using glutathione agarose beads followed by elution in the presence of free glutathione. The very mild elution conditions for release of fusion protein from the affinity matrix greatly minimise effects on its functional activity. The vector also contains a thrombin protease recognition site for cleaving *pogo* transposase from the fusion protein.

The ORF1-ORF2 and pGEX4T-2 DNA were digested separately with restriction enzymes *BamHI* and *XhoI*, ethanol precipitated, then ligated with each other for three hours at room temperature in the presence of T4 ligase. Competent *E.coli* NM522 cells were then transformed and allowed to grow overnight on selective L-Amp plates. The colonies containing the transposase coding sequence were selected by colony hybridisation as described in 2.2.2.16. The DNA of the positives was prepared and digested with *BamHI* and *XhoI*, to identify transformants containing the 1572 bp ORF1-ORF2 inserts (Fig. 3.2).

3.2.1.4 Checking the pGEX-*pogo* constructs by DNA sequencing

The positive constructs were examined by DNA sequencing of the junction of 5' end of the ORF1-ORF2 DNA and the pGEX4T-2 vector to confirm that the cloned DNA was in the correct reading frame for fusion protein expression. Since there is a hairpin structure at the C-terminus of the GST gene near the cloning site, a sequencing primer that has homology to the minus strand of the cloned gene was used to sequence the DNA at the junction region.

Fig. 3.3 shows the DNA sequence at the junction region of *pogo* ORF1-ORF2 with pGEX4T-2 vector by sequencing the minus strand using primer R39(-), as described in 2.1.8. Since the starting codon ATG for the first amino acid methionine of *pogo* transposase was omitted in order that the translation of the fusion protein starts from the GST, the three nucleic acid bases GGT encoding the second amino acid glycine of *pogo* transposase was immediately after the *BamHI* site GGATCC which encodes the last two amino acids of GST, indicating that *pogo* ORF1-ORF2 gene had

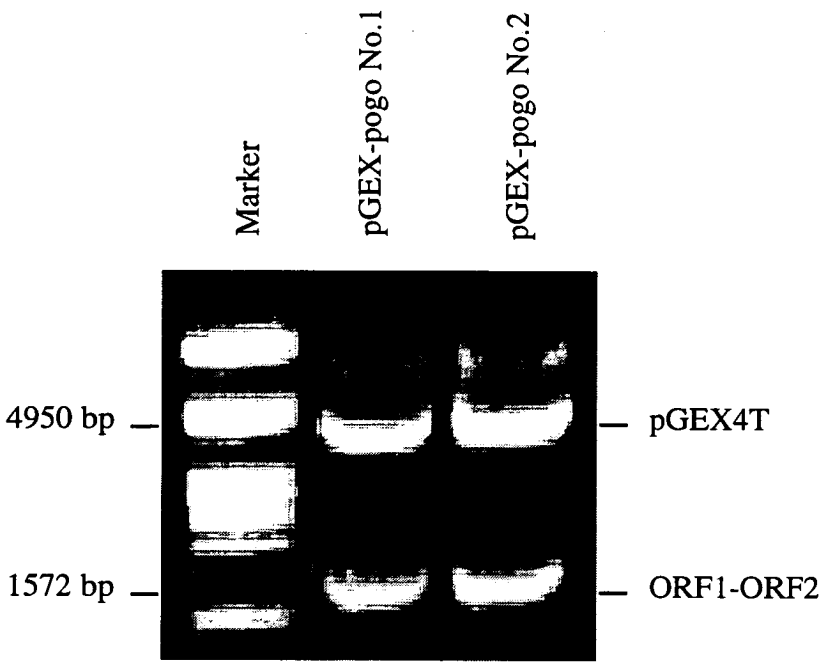


Fig. 3.2 Checking pGEX-pogo constructs by restriction digestion

pGEXpogo constructs were checked by restriction digestion with *BamHI* and *XhoI*. Two bands corresponding to pGEX(4950 bp) and pogo ORF1-ORF2 (1572 bp) respectively could be seen when the samples were run in a 1% agarose gel and visualised by ethidium bromide staining.

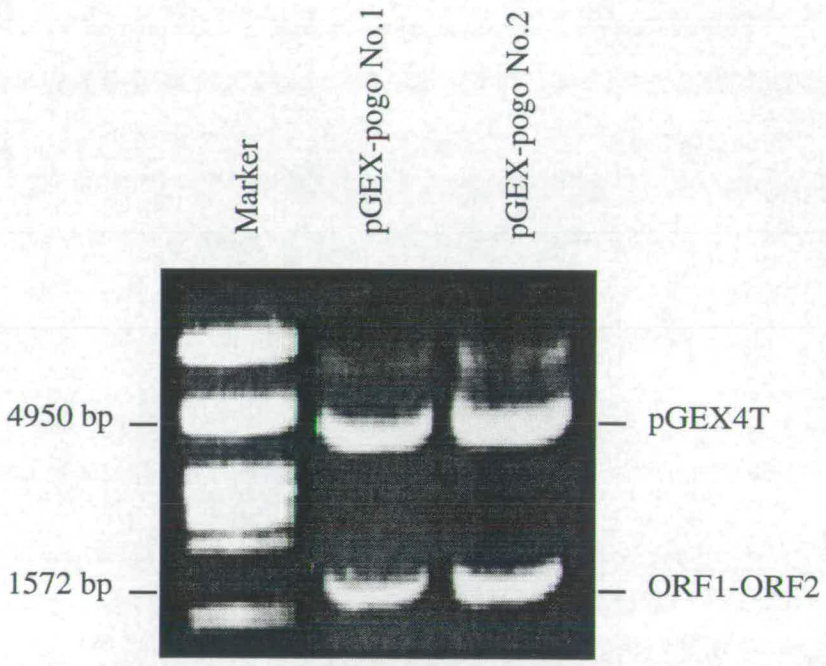
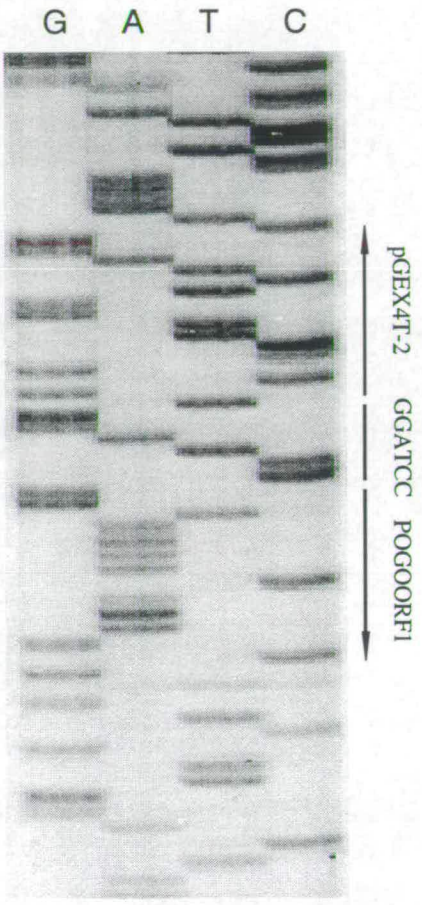


Fig. 3.2 Checking pGEX-pogo constructs by restriction digestion

pGEXpogo constructs were checked by restriction digestion with *BamHI* and *XhoI*. Two bands corresponding to pGEX(4950 bp) and pogo ORF1-ORF2 (1572 bp) respectively could be seen when the samples were run in a 1% agarose gel and visualised by ethidium bromide staining.



DNA sequence of the junction region:

CCAAAATCGGATCTGGTTCCGCGT**GGATCC**GGTAAAACAAAGCGTGTCGTTCTA

pGEX4T-2 sequence — *BamHI* site — pogo ORF1 sequence

Fig. 3.3 DNA sequence of the junction region of *pogo* ORF1-ORF2 with pGEX4T-2

The pGEXpogo consruct was checked by DNA sequencing. Since the first amino acid methionine of pogo transposase was omitted in order that the fusion protein was translated from the begining of GST, the codon GGT coding for the second amino acid of pogo transposase was immediately after the *BamHI* site GGATCC which codes for the last two amino acids of GST, indicating that the DNA was cloned into the correct reading frame for expression of pogo transposase as a GST fusion protein.

been cloned into the correct reading frame in the chosen vector, so the putative *pogo* transposase could be expressed as a fusion protein with GST.

3.2.2 Expression and purification of GST-transposase fusion protein

3.2.2.1 GST-transposase fusion protein was only expressed in pGEX-*pogo* transformed *E.coli* cells

Competent *E.coli* NM522 cells were transformed by the positive pGEX-*pogo* construct and 10 ml of overnight culture was grown from a single colony. The culture was then diluted 1:20 into fresh media and grown for 1 hr at 37 °C with vigorous shaking before protein expression was induced by IPTG for 3-7 hr. The culture was then harvested by centrifugation and resuspended in 1x PBS. The cells were lysed by sonication, and the insoluble fraction of the cell lysate was precipitated by centrifugation. The supernatant was removed and allowed to bind with glutathione agarose beads for 1 hr. After the binding, the beads were washed three times with 1x PBS. The soluble GST-transposase fusion protein, which bound to the glutathione agarose beads was visualised by SDS PAGE after being boiled off from the beads in boiling mix. NM522 cells and pGEX transformed NM522 cells were treated in the same way to produce control extracts.

Fig. 3.4 shows that GST-transposase fusion protein, which is 85.24 kD in size, was only present in pGEX-*pogo* transformed *E.coli* cells. It was specifically encoded by pGEX-*pogo*, as it was not present in either *E.coli* cells themselves or pGEX transformed *E.coli* cells.

3.2.2.2 The stability of GST-transposase fusion protein

Some of the fusion protein expressed in *E.coli*, especially eukaryotic proteins, are unstable and are degraded by proteases produced by *E.coli* cells. This was the case with the *pogo* GST-transposase fusion protein.

Fig 3.5 shows the protein expressed in pGEX-*pogo* transformed *E.coli* NM522 cells. Normally, two protein bands were seen. The larger one corresponded to the 85.2 kD GST-transposase fusion protein, while the smaller one represented the 27.6 kD GST. This could because the transposase part of some fusion protein did not fold properly therefore could be degraded by *E.coli* protease. In some cases only GST was visible because of a complete degradation of the transposase part. The DNA constructs were still positive for the latter cases as shown by restriction digestion. Moreover, the constructs show the same expression pattern even when different *E. coli* cells were used for expression.

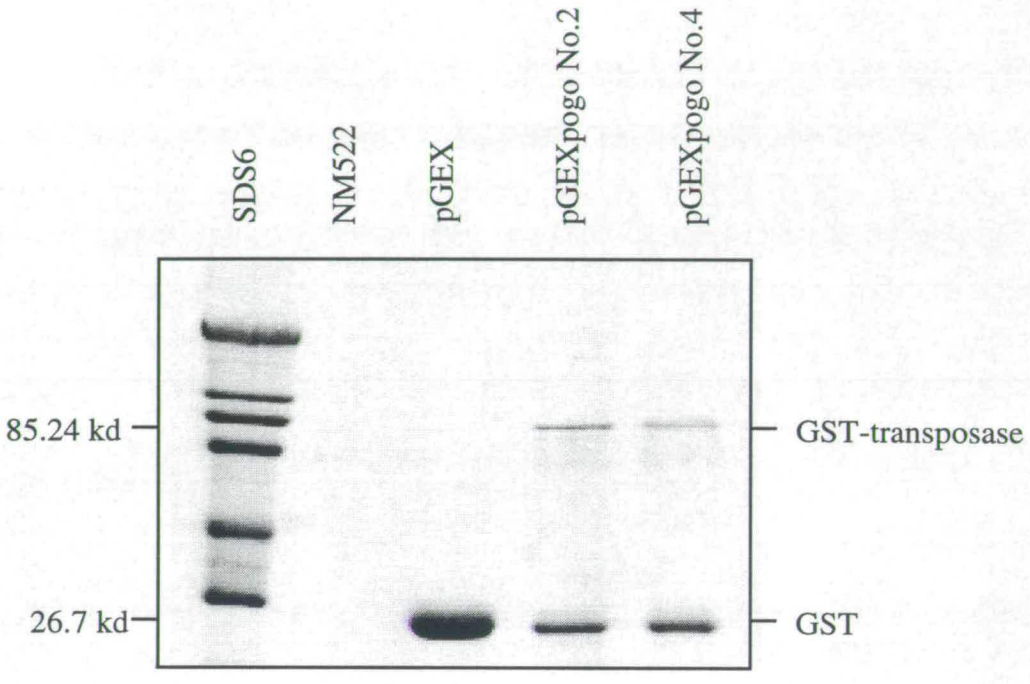


Fig. 3.4 Expression of GST-transposase fusion protein

The soluble proteins from NM522 cells, pGEX transformed NM522 cells and pGEXpogo transformeed NM522 cells were purified by using glutathion agarose beads, then run in a SDS PAGE. NM522 cells gave no bands; pGEX transformed NM522 cells gave one band corresponding to GST protein; pGEXpogo transformed NM522 cells gave two bands corresponding to GST-transposase fusion protein and GST respectively.
SDS6: SDS6 molecular weight marker.
NM522: proteins from E. coli NM522 cells.
pGEX: proteins from pGEX transformed E. coli cells.
pGEXpogo No.2: proteins from pGEXpogo No.2 transformed E. coli cells.
pGEXpogo No.4: proteins from pGEXpogo No.4 transformed E. coli cells.

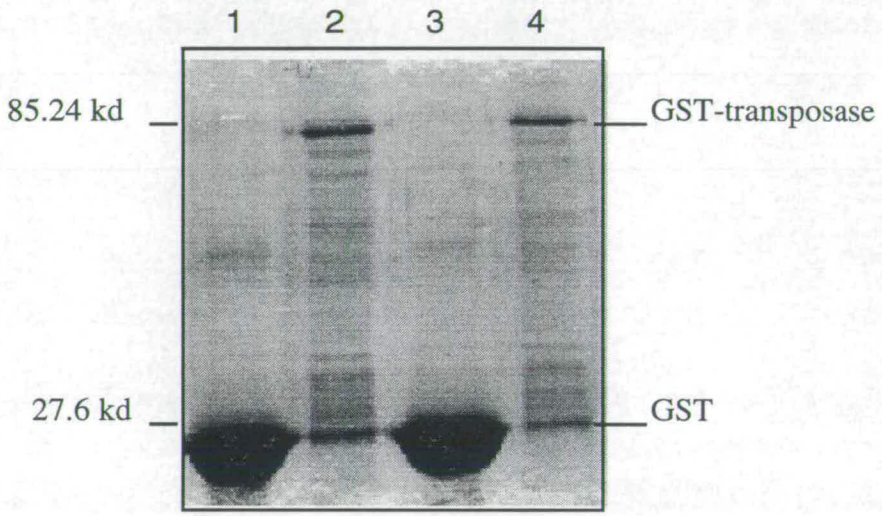


Fig. 3.5 GST-transposase fusion protein expressions of different pGEX-pogo constructs

Different positive pGEXpogo constructs (No.1-4) were checked for their GST-transposase expression by SDS PAGE. For constructs No.1 and No.3, only one band corresponding to GST could be seen; for constructs No.2 and No.4, two bands corresponding to GST-transposase fusion protein and GST respectively could be seen.

3.2.2.3 GST-transposase fusion is a partially soluble protein

Many proteins overexpressed in *E.coli* are insoluble or partially soluble. Since there was only a small amount of *pogo* GST-transposase fusion protein detected in the soluble fraction, the insoluble fraction, which was precipitated in the pellet after centrifugation of the cell lysate, was examined by boiling the insoluble pellet in boiling mix and visualised by SDS PAGE.

Fig. 3.6 shows the soluble *pogo* GST-transposase fusion protein which was in the supernatant and purified by binding with glutathione agarose beads, and the insoluble GST-transposase protein which was precipitated in the pellet after centrifugation. A strong band representing the 85.24 kD GST-transposase fusion protein was visualised in the insoluble fraction of the cell lysate, indicating some of the GST-transposase fusion protein was insoluble. Comparison of the amount of GST-transposase fusion protein in the soluble and insoluble fractions show that approximately 3/4 of the expressed protein was insoluble while only 1/4 of it was soluble.

3.2.2.4 Effects of temperature, IPTG concentration, length of induction time and presence of Triton-100 on the solubility of GST-transposase fusion protein

In order to increase the proportion of soluble GST-transposase fusion protein that could be used for activity studies, several factors which might effect the solubility of the protein were considered. The effects of temperature, IPTG concentration, length of induction time and presence of Triton-100 on the solubility of GST-transposase fusion protein were investigated.

Fig. 3.7 shows the overexpressed soluble GST-transposase fusion protein from the same wet weight of *E.coli* cells under various conditions. The cells produced approximately three times more soluble protein when the protein expression was induced at 30 °C than at 37 °C; changing the IPTG concentration from 0.1 mM to 1.0 mM had little effect on the amount of soluble protein. As the induction time changed from 3 to 5 hours, the soluble protein obtained from the same amount of cells decreased a bit, however the yield of soluble protein per unit volume of culture increased due to the increased cell density.

Fig. 3.8 shows the soluble protein obtained in and without the presence of 1% Triton-100 in the cell lysate. There was more soluble protein obtained in the presence of Triton-100 than without it.

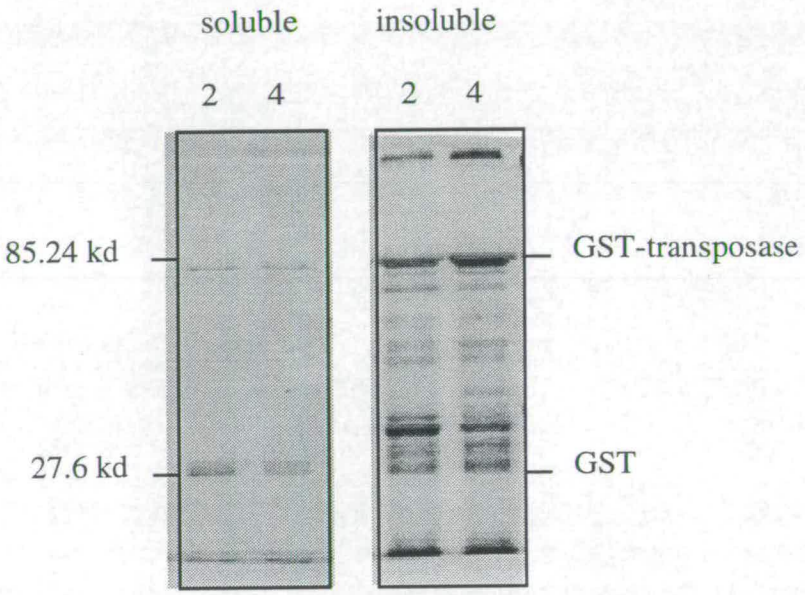


Fig. 3.6 Soluble and insoluble GST-transposase fusion protein expressed in *E.coli*

The soluble and insoluble GST-transposase fusion proteins from the same amount of *E.coli* cell were checked by SDS PAGE. Approximately1/4 of the protein is soluble while 3/4 of it is insoluble.
2: GST-transposase fusion protein from pGEXpogo No.2 transformed NM522 cells.
4: GST-transposase fusion protein from pGEXpogo No.4 transformed NM522 cells.

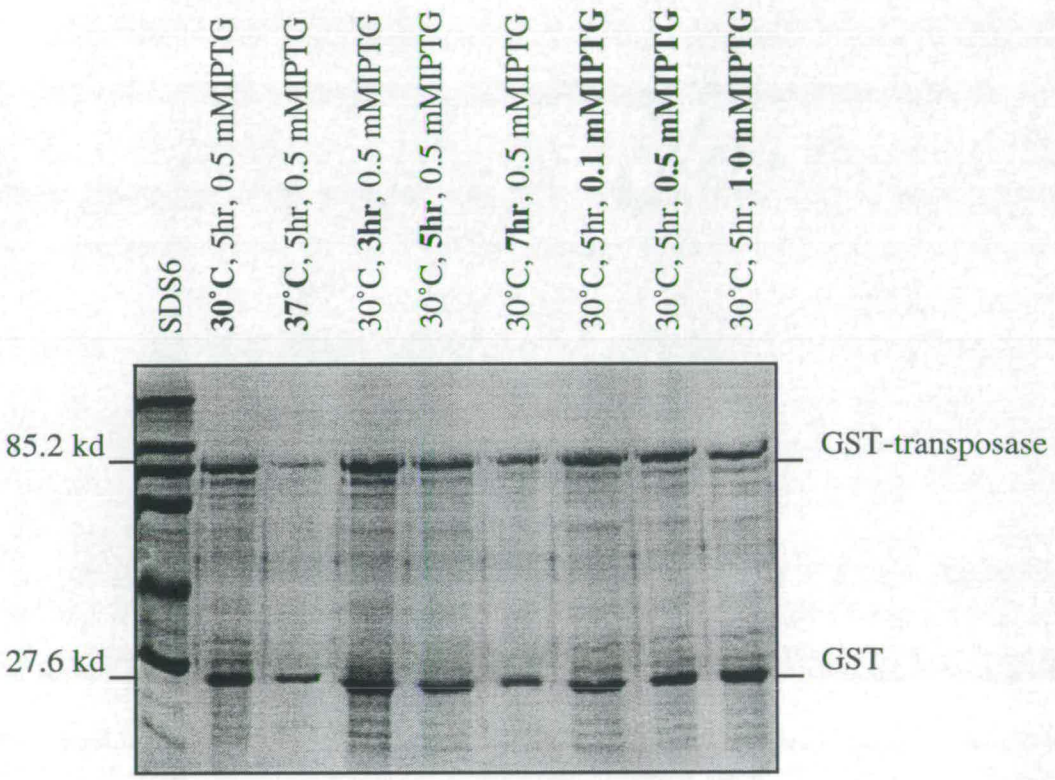


Fig. 3.7 Soluble GST-transposase fusion protein expression under various induction conditions

The soluble GST-transposase fusion protein from same wet weight of pG EXpogo transformed NM522 cells induced under different temperature (30°C and 37 °C), induction time (3 hr, 5 hr and 7 hr) and IPTG concentration (0.1 mM, 0.5 mM and 1.0 mM) were run in a SDS PAGE and visualised by coomassie staining, so that the relative amount of each sample could be seen and compared with the others.

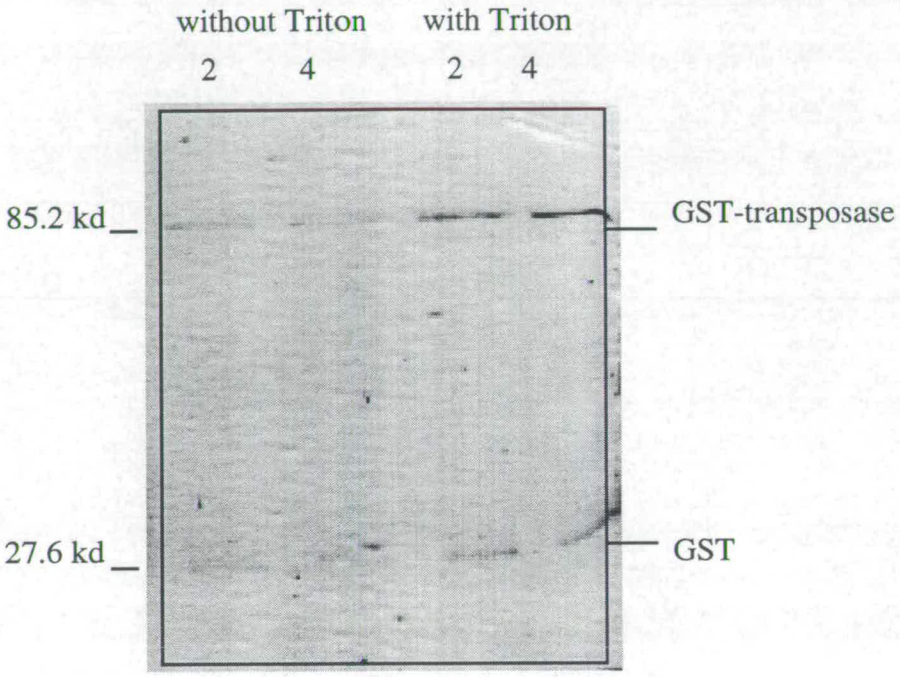


Fig. 3.8 Soluble GST-transposase fusion protein obtained without and with the presence of Triton-100

The soluble GST-transposase fusion protein obtained from the same wet weight of pGEXpogo transformed E.coli cells in the presence and absence of 1% Triton-100 were run in a SDS PAGE and visualised by coomassie staining, so that the relative amount of each sample could be seen and compared with the others.
2: soluble GST-pogo protein from pGEXpogo No.2 transformed *E.coli* NM522 cells.
4:soluble GST-pogo protein from pGEXpogo No.4 transformed *E.coli* NM522 cells.

3.2.3 Cleavage of GST tag from GST-transposase fusion protein

GST, as a tag of the expressed protein, greatly facilitates the expression and purification of the fusion protein. But GST itself , as a 27.6 kD protein at the N-terminus of the expressed protein, might affect the conformation and/or the activity of the protein. So it is probably ideal to remove the GST tag from the fusion protein in some studies. The thrombin cleavage site, which is very close to the N-terminus of the expressed protein in a pGEX4T vector, makes it possible to remove the GST tag by site specific protease digestion.

The GST-transposase fusion protein from the cell lysate was bound to glutathione agarose beads and then cleaved by digestion with thrombin. The released transposase protein could be easily separated from the GST and the un-cleaved fusion protein, which was still bound with the affinity matrix.

Fig 3.9 shows the transposase as a GST fusion protein when it was eluted from the agarose beads by solution containing reduced glutathione, and as a free protein when it was cleaved by thrombin from the GST tag.

3.3 Discussion

The *E. coli* system was used for expression of *pogo* transposase in this thesis, because the techniques necessary are relatively simple, and the time need is short. Also it is cheap to grow and can get relatively large amount of expression. Although the overexpressed protein was only partially soluble, sufficient GST-transposase was produced to provide enough protein for the studies.

Good expression of a protein encoded by a foreign gene in *E. coli* can be obtained by making fusions of the cloned gene to a highly expressed carrier protein which provides the necessary signal for good expression. The carrier sequence can also code for an entire protein which can be exploited in purifying the protein. GST fusion is one of the examples. By taking the advantage of using the GST carrier, *pogo* transposase has been expressed as a GST fusion protein in this thesis, and a reasonable amount of the GST-transposase fusion protein was produced. The soluble fraction of the protein was purified from lysed cells under nondenaturing conditions by absorption with glutathione-agarose beads, followed by elution in the presence of free glutathione.

Many proteins form insoluble inclusion bodies when expressed at high level in *E. coli* (Schein, 1989). Several factors, such as temperature, expression level, the genotype of *E.coli* strains and the carrier protein, are supposed to affect the solubility of expressed fusion proteins (Bishai et al., 1987; Schein, 1989; P. Riggs, 1991). For the *pogo* GST-transposase fusion protein, decreasing induction temperature from 37°C

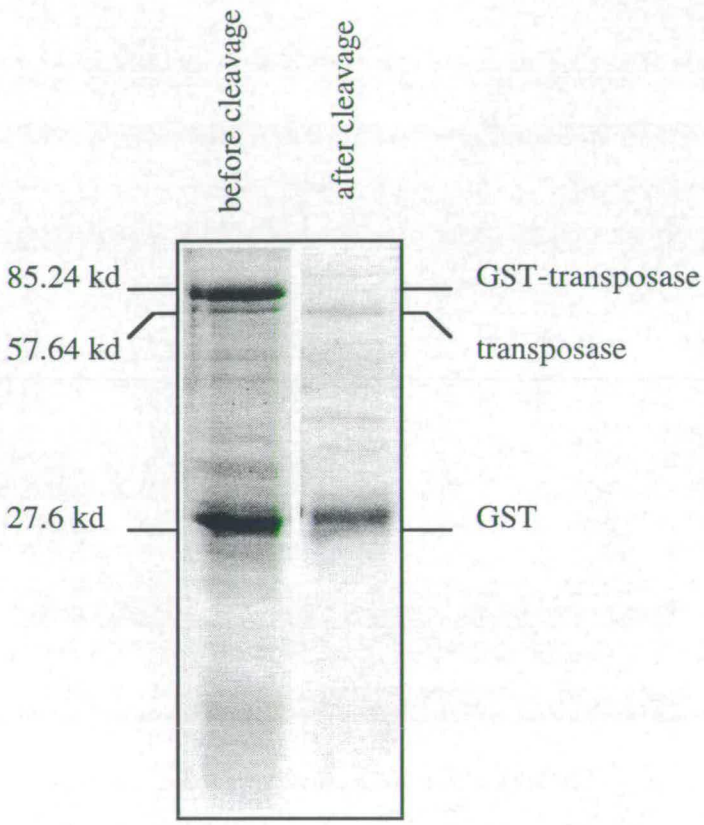


Fig. 3.9 Cleavage of pogo transposase from GST-transposase fusion protein

The GST-transposase fusion protein from cell lysate was purified by binding to glutathion agarose beads and then eluted by solutions containing reduced glutathion, or cleaved by digestion with thrombin. The samples before and after thrombin cleavage were run in an SDS PAGE and visualised by coomassie staining.

to 30°C did increase the amount of soluble protein formation. When the *E.coli* cell was changed from NM522 to BL21 DE3 plyS, there was little affect on the amount of soluble protein produced (data not shown). Different vectors were tried by cloning the transposase gene into a pRSET vector to express the *pogo* transposase as a fusion with a six histidine tag, but little expression of the protein was obtained (data not shown); however, when pGEX4T-2 vector was used, a large amount of GST-transposase fusion protein was expressed. Another way to alter the solubility of the expressed protein is the addition of mild detergent after cell lysis. The detergent is thought to decrease the interaction of expressed protein with *E.coli* proteins and prevent its precipitation with insoluble *E.coli* proteins. This increases the yield of the soluble protein. 1% Triton was used in this thesis and the solubility of GST-transposase fusion protein did obviously increase.

Foreign proteins, especially eukaryotic proteins, are often unstable when they are expressed in *E. coli*. The carrier protein can sometimes stabilise an expression fusion protein (Lee et al., 1984). Sometimes the expressed protein is degraded but the carrier protein is not. Moreover, fusion proteins are sometimes cleaved *in vivo* at the fusion joint between the carrier and the expressed portion of the fusion. These facts about fusion proteins are consistent with a model in which the carrier and the rest of the protein form independent domains. According to this theory, the carrier domain may fold correctly while the expressed protein does not and is degraded. There may also be cases where both domains fold correctly but the joint region between them is sensitive to one or more *E. coli* proteases. For the GST-transposase fusion protein in this thesis, this model can explain why free GST was seen by SDS PAGE in addition to the GST-transposase fusion protein, and why only GST was visible in some samples. This is probably because the *pogo* transposase domain of some fusion protein didn't fold properly therefore had been partially or completely degraded by some proteases from *E.coli*, only the properly folded GST domain was purified by binding with glutathione agarose beads and seen by SDS PAGE.

When fusion proteins produced in *E. coli* are used in activity studies, it is frequently desirable and sometimes necessary to separate the carrier domain from the expressed proteins. For foreign DNA expressed in pGEX2T or pGEX4T, the carrier protein and the protein of interest can be seperated by site-specific protease cleavage of the amino acid sequence between the GST carrier and the foreign polypeptide by thrombin (Gearing, et al., 1989). Two alternatives are available for cleavage of fusion proteins: to conduct the cleavage reaction on the pure fusion protein solution (Smith and John, 1988) or to cleave the foreign polypeptide from GST while it is still bound to glutathione-agarose beads (Garing et al., 1989). The second procedure is faster and

faster and more efficient, and has been used for cleavage the GST carrier from GST-transposase fusion protein in this thesis.

Chapter 4

THE SEQUENCE SPECIFIC DNA BINDING ACTIVITY OF *POGO* TRANSPOSASE

4.1 Introduction

It is transposases that are encoded by transposable elements themselves that are responsible for the replicative spread of transposable elements within and some times between large genomes. As an initial step of the transposition process, a transposase must first recognise and interact site-specifically with the ends of the element to form a highly organised transposition complex. The transposase can then catalyse the endonuclease cleavage and strand transfer reactions that follow. Specific DNA binding ability is therefore an essential character of every transposase.

The experiments described in this chapter were designed to determine whether *pogo* transposase made in *E. coli* has the DNA binding activity expected of the native enzyme. The DNA binding activity of a protein can be determined by a gel retardation assay. In this assay, DNA is radioactively labelled, so that its movement in a gel can be easily followed by autoradiography. When a protein binds to the DNA, the DNA-protein complex usually moves more slowly in a non-denaturing polyacrylamide gel, resulting in a retarded band compared to that of free DNA probe. The specificity of the DNA-protein binding can be demonstrated by a competition experiment. In this experiment, increasing amounts of unlabelled specific or non-specific competitors are added to the binding reaction to compete with the radioactively labelled probe for binding with the protein. The specific competitor is usually the unlabelled probe DNA, while the non-specific competitor can be any DNA of different sequence to the probe, but of similar size. Non-specific DNA-protein binding can be easily titrated away by the addition of non-specific competitor, while the specific binding can't. The latter can only be competed away by a specific competitor since the specific competitor has much higher affinity to the protein than the non-specific competitors.

Transposases are usually thought to bind specifically with similar sequences at both end of the element. The inverted repeats, which are symmetrically located at both ends of the element, are more likely to be the candidates for this job. This has been show to be the cases for *Tc1* and *Tc3* elements (Vos et al., 1993; Colloms et al., 1994). The transposases of these elements bind specifically to DNA sequences within their inverted repeats. However it's not true for *P* and *Ac* elements, their transposases have been show to bind specifically to the subterminal sequences of these elements (Kaufman et al., 1989; Kunze & Starlinger, 1989). For the *pogo* element, as it was not clear which was going to be the case at this stage, a 131 bp sequence from 5' end of the element, which including the 21 bp inverted repeat and the 110 bp subterminal sequence, was used as the probe in the experiments described in this chapter in order to cover all the possibilities. The results of these experiments show that *pogo* transposase binds to sequences in this region in a sequence-specific manner.

4.2 Results

4.2.1 GST-transposase fusion protein is responsible for binding to the transposon end sequence

GST-transposase fusion protein overexpressed in pGEX-*pogo* transformed *E.coli* cells was purified by binding to glutathione agarose beads and then eluted from the beads with reduced glutathione. The purified protein was allowed to bind to the radioactively labelled DNA probe, in the presence of non-specific competitor poly dI-dC. The samples were then run in a non-denaturing polyacrylamid gel, and the mobility of the labelled DNA was visualised by autoradiography. Protein from untransformed *E.coli* cells and cells transformed with pGEX were treated in the same way, and the results were compared.

Fig. 4.1 shows that the DNA probe only bound to the GST-transposase fusion protein purified from pGEX-*pogo* transformed *E.coli* cells. It didn't bind to either the proteins from *E.coli* cells or GST protein from pGEX transformed *E. coli* cells. This suggests that GST-transposase fusion protein is responsible for the specific binding to the transposon end sequence.

4.2.2 The binding of GST-transposase fusion protein to the transposon end is specific binding

A competition experiment was conducted to further investigate the sequence specific DNA binding activity of GST-transposase fusion protein. Unlabelled probe DNA was used as specific competitor, while the first 190 bp DNA sequence of *mariner* element, which is of different sequence to that of *pogo*, was used as non-specific competitor (Fig. 4.2). The binding reactions of GST-transposase fusion protein with the labelled probe were carried out in the presence of poly dI-dC plus either specific or non-specific competitors, and the results were compared.

Fig. 4.2 shows that the DNA-protein complex of the probe with GST-transposase fusion protein could be titrated away by a 100 fold molar excess of specific competitor, but not by the same molar excess of non-specific competitor. This indicates that GST-transposase fusion protein has much higher affinity for the 5' end of the *pogo* transposon than for other DNA sequences.

4.2.3 Free *pogo* transposase can bind specifically to the transposon end sequence

To further test that the specific DNA binding activity of GST-transposase fusion protein was due to the presence of *pogo* transposase, rather than the GST tag,

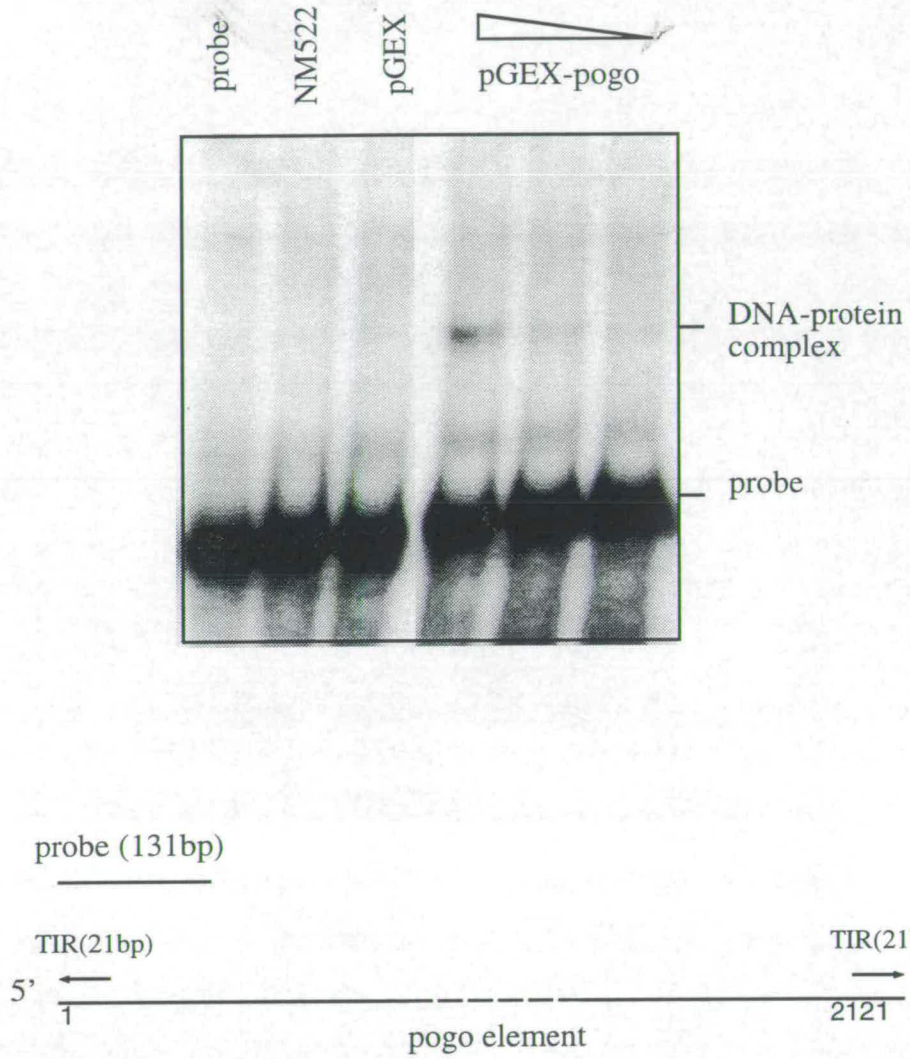
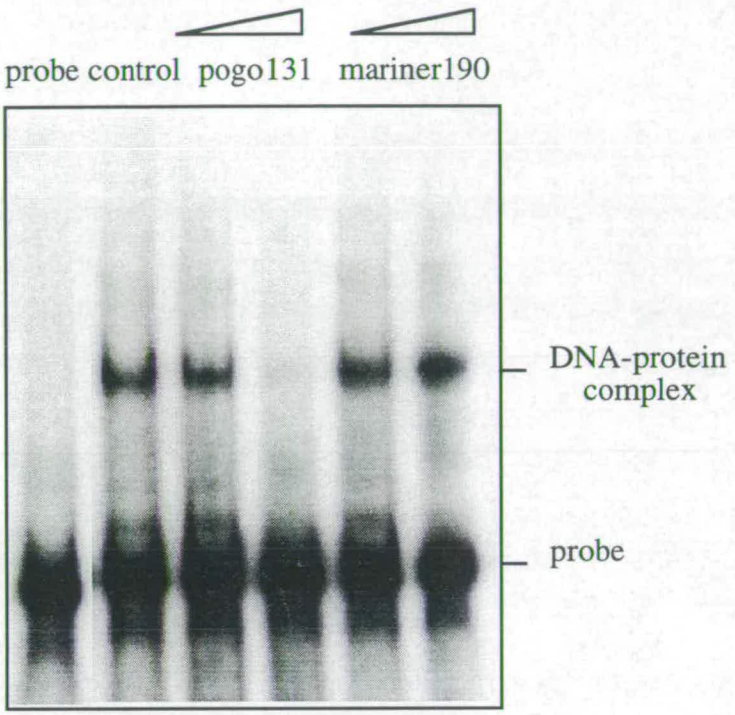


Fig. 4.1 Specific binding of GST-transposase fusion protein to the transposon end sequence

Probe: the first 131 bp DNA fragment of pogo element was end labelled and used as probe.
NM522: protein from E.coli NM522 cells (~ 0.02 g wet weight).
pGEX: GST protein from pGEX transformed E.coli cells (~0.02g wet weight).
pGEX-pogo: GST-transposase fusion protein from pGEX-pogo transformed E. coli cells (0.02, 0.01, and 0.005 g wet weight respectively).
1 ng of probe DNA was used per reaction.
1µg of poly dI-dC was used per reaction as non-specific competitor.



mariner 1-190 sequence

CCAGGTGTAC AAGTAGGGAA TGTCGGTTCG AACATATAGA
TGTCTCGCAA ACGTAAATAT TTATCGATTG TCATAAAACT TTGACCTTGT
GAAGTGTCAA CCTTGACTGT CGAACCACCA TAGTTTGGCG CGAATTGAGC
GTCATAATTG TTTACTCTCA GTGCAGTCAA CATGTCGAGT TTCGTGCCGA

pogo 1-131 sequence

CAGTATAATT CGCTTAGCTG CATCGATAGT TAGCTGCATC GGCAAGATA
CTGCATTATT TTTCCATTTT TTTGTGTGAA TAGAAAATTT GTACGAAAA
TCATACGTTT GCTGCATCGC AGATAACAGC C

Fig. 4.2 Binding of GST-transposase fusion protein to the transposon end sequence competed by specific and non-specific competitors

probe: 1-131 of pogo sequence
control: neither pogo131 nor mariner 190 was added as competitor
pogo131: 10-100 fold molar excess of pogo 1-131 was added as specific competitor
mariner 190: 10-100 fold molar excess of mariner 1-190 was added as non-specific competitor
1 ng of probe DNA was used per reaction.
1 μg of poly dI-dC was used per reaction as non-specific competitor.
~3 μg of protein per binding reaction.

the DNA binding activity of the free transposase was studied. Free *pogo* transposase, which was obtained by thrombin cleavage of the transposase from GST-transposase fusion protein, was allowed to bind to the DNA probe in gel retardation assay, in the presence of non-specific competitor poly dI-dC, and the result was compared with that of GST-transposase fusion protein.

Fig. 4.3 shows that free *pogo* transposase could bind specifically to the probe, and form a characteristic complex when run in a polyacrylamide gel. The DNA-protein complex formed by the free protein moves faster than that of the fusion protein due to the smaller size of the free protein. Furthermore, Fig. 4.1 shows that the GST itself has no specific binding ability with the same DNA probe. These results demonstrate that *pogo* transposase, rather than the GST tag, was responsible for the specific binding of the fusion protein to the *pogo* end sequence.

4.3 Discussion

The interaction of a protein with a specific DNA sequence is best analysed in the presence of non-specific competitor DNA. This competitor should not contain any copies of the specific binding site for the protein. A non-specific DNA of known sequence or unnatural polynucleotides with simple repeat sequences such as poly d(I-C) are normally used. On addition of relatively large amount of non-specific competitor DNA, all non-specific complexes with the labelled DNA should be titrated away. Any complexes that remain represent the specific DNA-protein association. This can be confirmed by the disappearance of the complex when titrated with unlabelled substrate DNA, the specific competitor. In gel retardation assays in this thesis, 1.0 μg of poly d(I-C) was used as non-specific competitor in each reaction. In the competition experiment in this chapter, an additional 100 fold molar excess of the first 190 bp sequence from *mariner* element was used as non-specific competitor, but still failed to compete away the DNA-protein complex. The complex could only be competed away by the same molar excess of unlabelled probe DNA. This suggests that the interaction of *pogo* transposase with the transposon end sequence is a specific reaction.

It was a concern that the GST, as a 27.6 kD tag at the N-terminus of *pogo* transposase, might interfere with the activities of the transposase. However, results from this chapter show that the GST-transposase fusion protein can bind specifically to the transposon end sequence; This binding is due to the transposase, rather than the GST tag or other proteins, and the GST-transposase fusion protein can interact as well with the transposon end sequence as the free transposase does. These data suggest that the GST tag and the transposase form two independent domains within the fusion protein. The GST domain might have little effect on the transposase domain, or at least

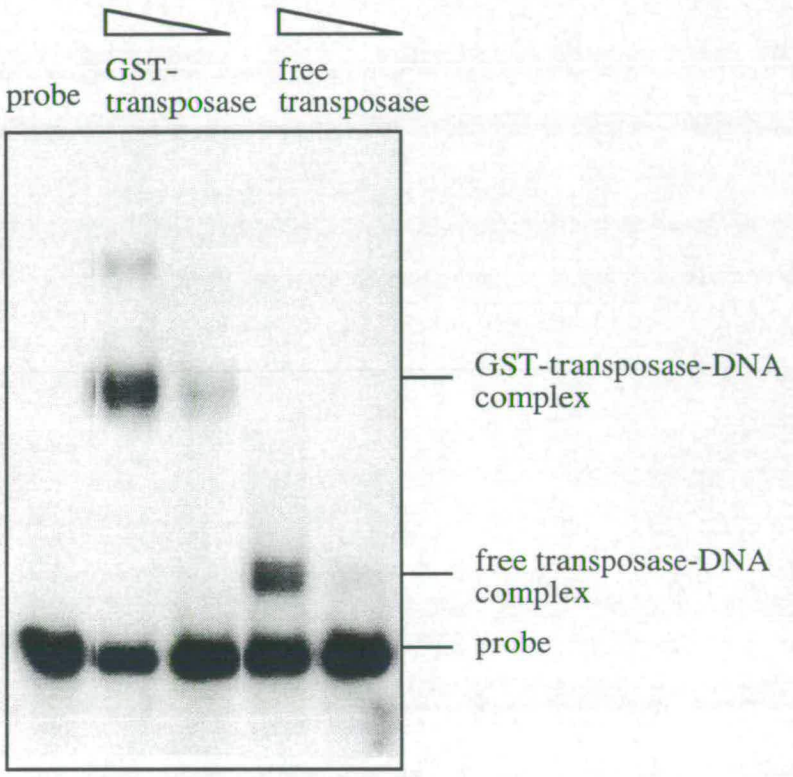


Fig. 4.3 Specific bindings of GST-transposase fusion protein and free transposase to the transposon end sequence

probe: 1-131 of pogo sequence
GST-transposase: GST-transposase fusion protein (~3 & 0.5 μg)
free transposase: transposase protein without GST tag (~same molar as fusion protein
1 ng of probe DNA was used per reaction.
1 μg of poly dI-dC was used per reaction as non-specific competitor.

on the DNA binding motif of the transposase, so the transposase domain can still interact specifically with the transposon end sequence. Furthermore, these results also suggest that it is possible to use the GST-transposase fusion protein to study the DNA binding activities of the transposase.

Chapter 5

THE TRANSPOSASE BINDING SITES OF *POGO* ELEMENT

5.1 Introduction

As an initial step of transposition, the transposase has to recognise and bind to both ends of the transposon to form a synaptic transposition complex, before it catalyses the subsequent endonuclease cleavage and strand transfer reactions. The experiments described in chapter 4 show that *pogo* transposase can bind sequence specifically to the first 131 bp sequence at the 5' end of the element. The experiments in this chapter were designed to investigate these further to determine the shortest DNA sequence (transposase binding site) that is involved in the specific interaction with the transposase, and also whether the transposase interacts similarly with both ends of *pogo*.

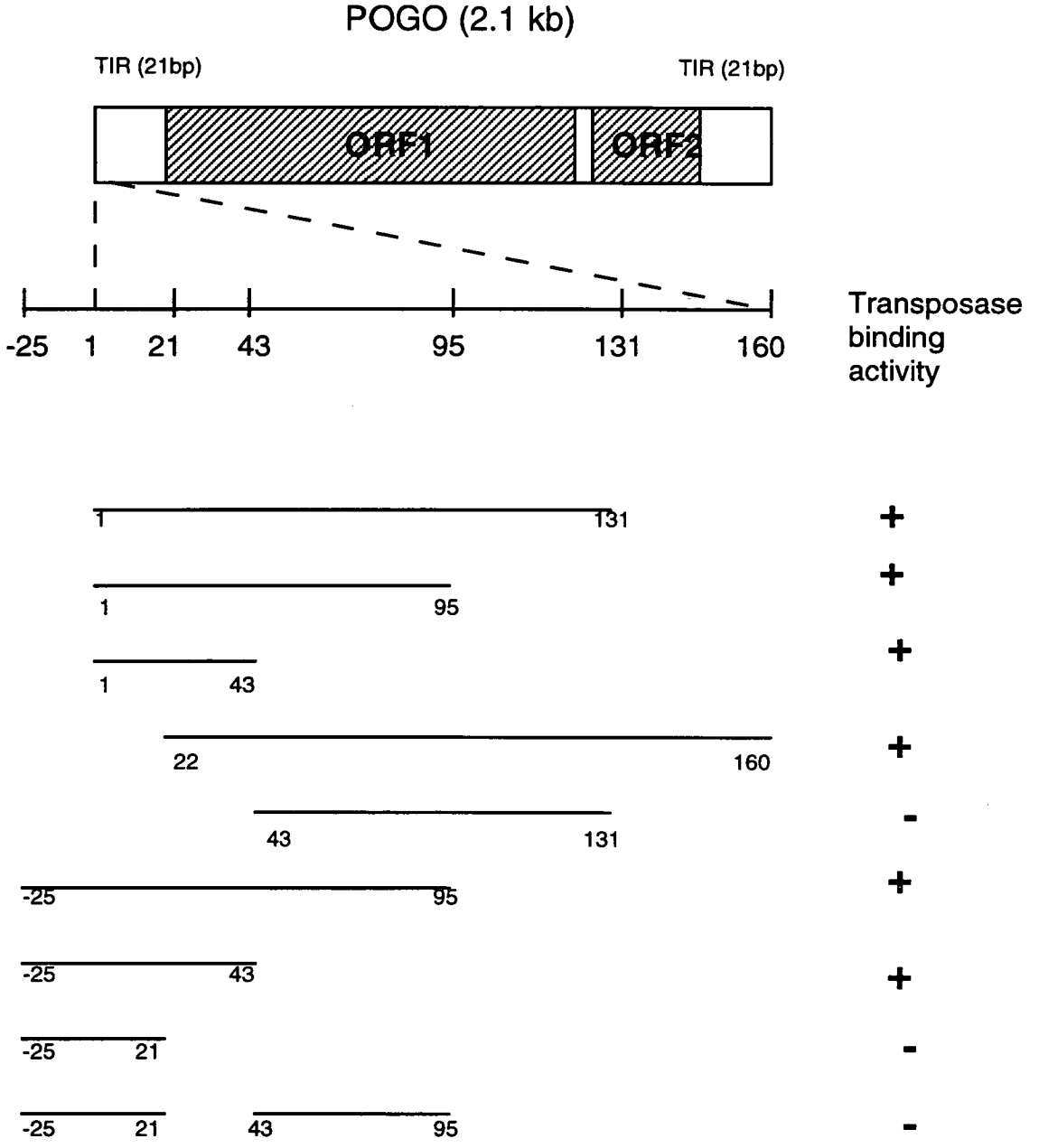
5.2 Results

5.2.1 Mapping transposase binding sequence at the 5' end of *pogo* element

To isolate the transposase binding sequence at the 5' end of *pogo* element, three groups of terminally deleted DNAs were made from the 131 bp probe used in chapter 4, which contains the 21 bp terminal inverted repeat and 110 bp subterminal sequence from 5' end of *pogo*. Each of these deleted DNAs was used as a probe and its ability to bind to the transposase was tested in gel retardation assay. A 22 bp sequence was finally isolated and shown to be responsible for the specific binding to the transposase (Fig. 5.1).

The first group of probes was made by deleting the 131 probe from 3' end to 95 bp then to 43 bp, to test whether the DNA sequence near the 5' end is involved in the specific binding to the transposase. Fig. 5.2 shows that both the 95 bp and 43 bp probes could bind specifically to the transposase. These indicate that the remaining DNA sequence 1-43 contains the transposase binding site. The transposase bound less efficiently to the 43 bp probe than to the 95 bp probe, probably because the transposase binding site is too close to one of the ends of the 43 bp probe, there is no sufficient sequences for holding the transposase properly.

The second group of probes was made by deleting the 131 probe from the 5' end by the first 21 bp and 42 bp sequence respectively, to test whether the 21 bp inverted repeat or the 22-43 sequence is responsible for the binding to the transposase. Fig. 5.3 shows that when the 21 bp inverted repeat was deleted from the 5' end of the 131 bp probe which was extended from the 3' end until 160 bp, the probe could bind to the transposase; when the first 42 bp of the 131 probe was deleted, the probe failed to bind to the transposase. These data indicate that nucleotide sequence 22-43, rather



5.1 Mapping pogo transposase binding site by deletions

Three groups of teminally deleted probe DNA were made and tested for their ability to bind with the transposase.

The first group was deleted from the 3' of the 131 probe. The second group was deleted from the 5' of the 131 probe. The third group was first extended the 131 probe to the left hand side for 25 bp, then deleted from the 3' end. A final probe was made by deleting the transposase binding sequence 22-43 identified by the above deletions, to test the transposase binding ability of the flanking DNA.

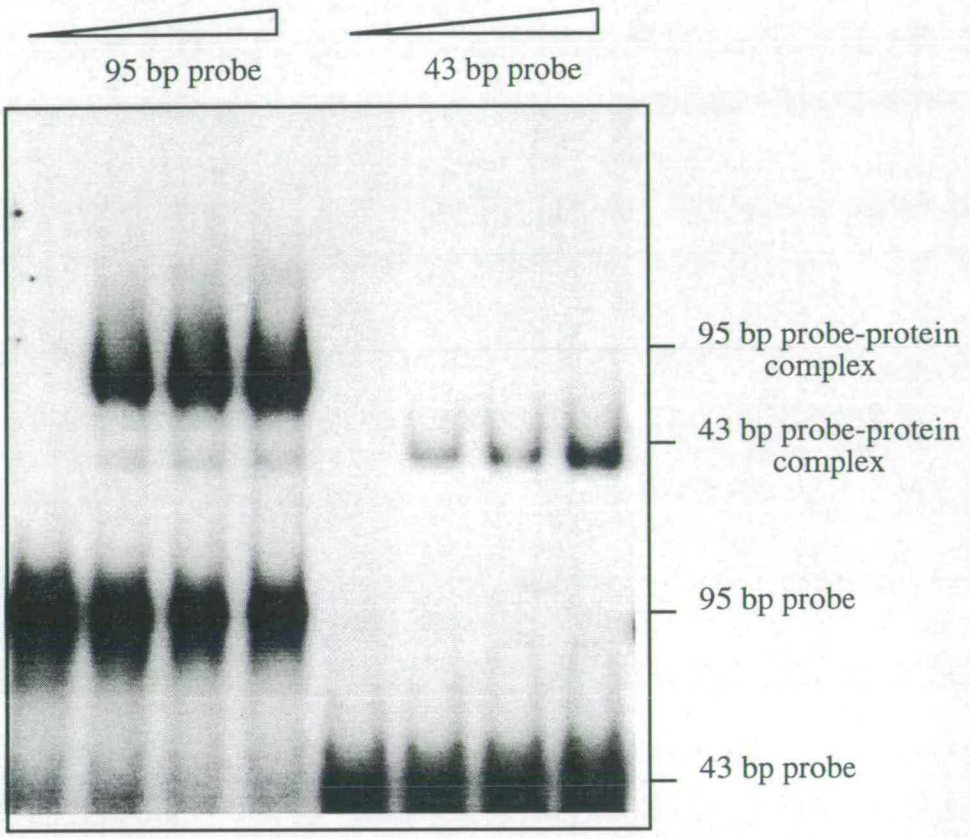


Fig. 5.2 The specific binding of pogo transposase to the 95 bp and 43 bp probe

The first 131 bp sequence of pogo element was deleted from the 3' end to 95 bp, and then to 43 bp. The transposase binding activities of these deleted probes were tested by gel retardation assay.
95 bp probe: 1-95 bp of pogo sequence.
43 bp probe: 1-43 bp of pogo sequence.
0, 5, 10 and 15 µl of pogo transposase N75 (~ 0.2 µg/µl) were used for binding with each probe.
~1 ng of probe DNA was used for each reaction.

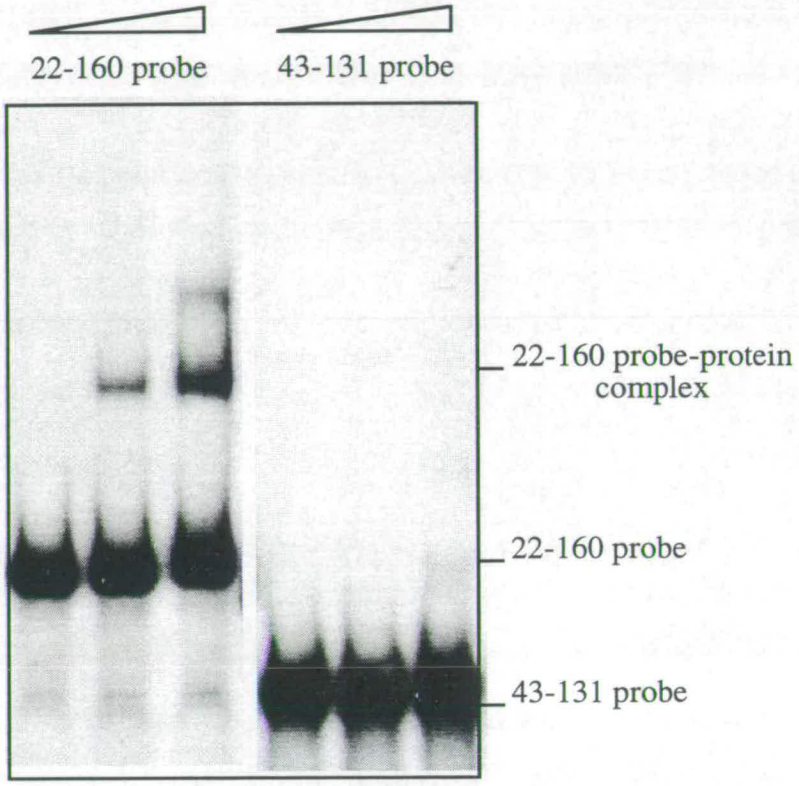


Fig. 5.3 Specific binding of pogo transposase to the 22-160 and 43-131 probes

The first 21 bp inverted repeat sequence and the first 43 bp sequence were deleted from the 5' end of 160 bp and 131 bp probes respectively. The deleted probes were tested for their abilities to bind to the transposase by gel retardation assay.
22-160 probe: 22-160 bp of pogo sequence.
43-131 probe: 43-131 bp of pogo sequence.
0, 5 and 15 µl of pogo transposase N75 (~ 0.2 µg/µl) were used for binding to each probe.
~1 ng of probe DNA was used for each reaction.

than the 21 bp terminal inverted repeat, contains the transposase binding site, which is involved in the specific binding to the transposase.

The third group of probes was made by extending the 131 probe to the left hand side for 25 bp, then deleting from the 3' end of the probe to 95 bp, 43 bp, and 21 bp respectively. The 25 bp left hand extension was made in order to provide more holding sequences for the transposase, in case the transposase binding site is too close to the 5' end of the probe. Fig. 5.4 shows that with the additional 25 bp extension to the left hand side from the 5' end, the first 95 bp and the first 43 bp probes could bind to the transposase, while the first 21 bp probe, which is the terminal inverted repeat of the element, could not bind to the transposase. These results further indicate that the transposase binding site is located within nucleotide sequence 22-43, rather than the first 21 bp inverted repeat sequence. The transposase binds less efficiently to the 43 bp probe than to the 95 bp probe even though there are 25 bp additional sequence at the 5' end, suggesting the transposase binding site might be located near the 3' end of the 43 bp probe.

A final probe was made by deleting the sequence 22-43 bp which was shown to be the transposase binding sequence from the previous deletions, to test the transposase binding ability of the flanking DNA sequence (Fig. 5.1).

Fig. 5.5 shows that when the DNA sequence 22-43 was removed from the 95 bp probe having the 25 bp left hand extension, the probe made from the remaining flanking sequence failed to bind to the transposase. This further indicates that the 22-43 sequence, rather than the flanking DNA, is involved in the specific binding to *pogo* transposase.

5.2.2 The transposase binding region at the 3' end of *pogo* element

A transposase is supposed to bind to both ends of the element. The transposase binding site at the 5' end of the *pogo* element has been shown to be located within the first 22-43 bp sequence, rather than the 21 bp terminal inverted repeat. Since the inverted repeat sequences are the same at both ends of the element, the transposase is likely to bind to a subterminal sequence at the 3' end of the element. To test this, a DNA fragment from nucleotide 1960-2100 (3'no-IR), which is the sequence immediately inside the 21 bp inverted repeat at the 3' end of the element, was used as probe. The transposase binding activity of this probe was analysed and compared with that of the 5' subterminal sequence 22-160 (5'no-IR).

Fig. 5.6 shows that the 3'no-IR probe could bind specifically to the transposase. This indicates that the subterminal sequence 1960-2100 at the 3' end of *pogo* element contains the specific transposase binding site.

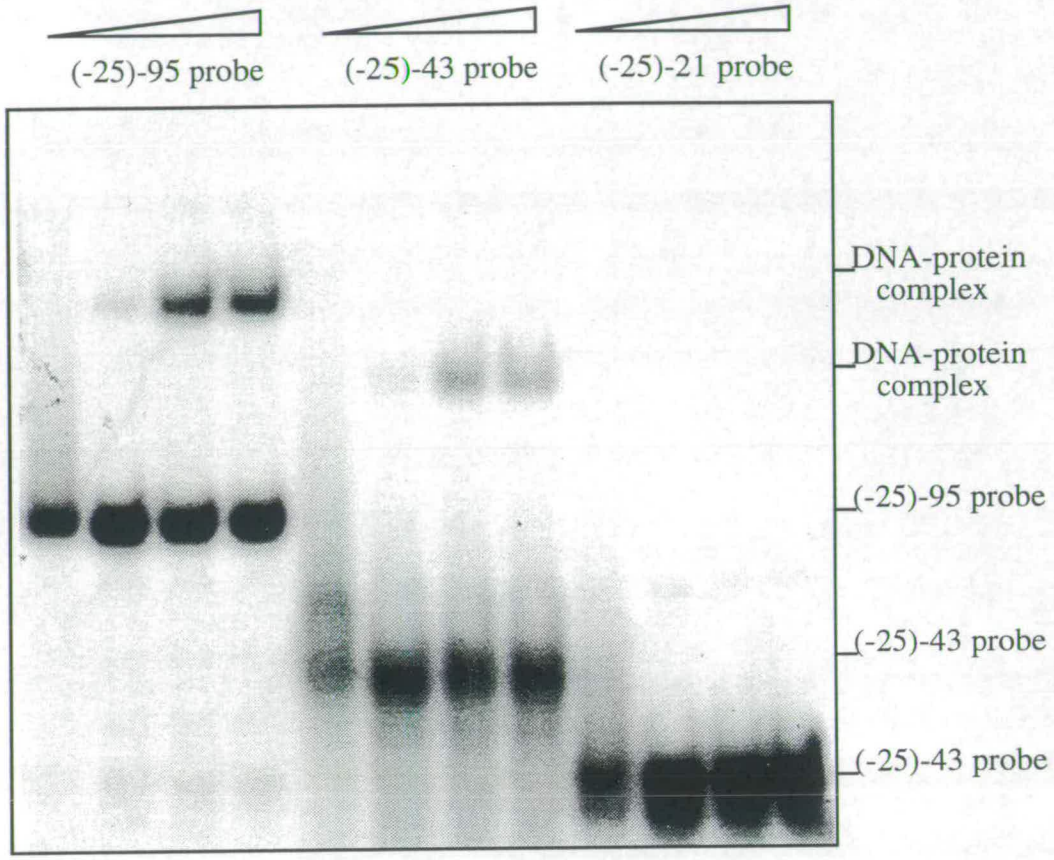


Fig. 5.4 Specific binding of pogo transposase to the (-25)-95, (-25)-43 and (-25)-21 probes

The first 131 bp sequence of pogo was extended from 5' end for 25 bp, then deleted from 3' end to 95 bp, 43 bp and 21 bp. The transposase binding activities of these deleted probes were tested by gel retardation assay.
(-25)-95 probe: pogo 1-95 sequence with additional 25 bp to the left hand side.
(-25)-43 probe: pogo 1-43 sequence with additional 25 bp to the lelt hand side.
(-25)-21 probe: pogo 1-21 sequence with additional 25 bp to the left hand side.
0, 1, 6 and 10 µl of pogo transposase N75 (0.2 µg/µl) were used for binding to each probe.
~1 ng of probe DNA was used for each reaction.

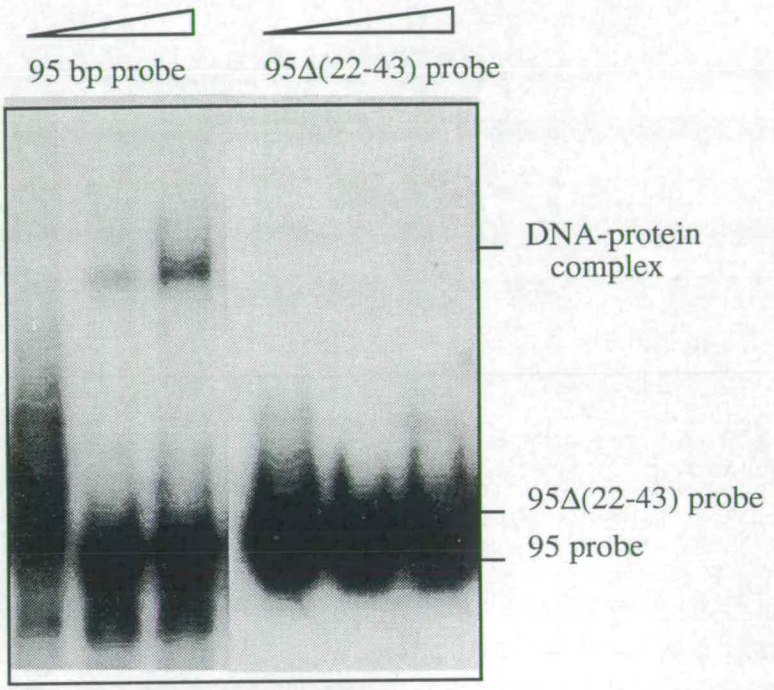


Fig. 5.5 The specific binding of pogo transposase to the 95 probe and the 22-43 internally deleted 95 probe

The DNA sequence 22-43 was deleted from the first 95 bp of pogo sequence with 25 bp left hand extension. The transposase binding activities of this deleted probe was analysed by gel retardation assay, and compared with that of the 95 bp probe.
95 bp probe: 1-95 bp of pogo sequence.
95Δ(22-43) probe: 1-95 of pogo sequence, with extended 25 bp to the left handside, and deleted from 22-43 sequence.
0, 1 and 6 µl of pogo transposase N75 (~0.2 µg/µl) were used for binding to each probe.
~1 ng of probe DNA was used for each reaction.

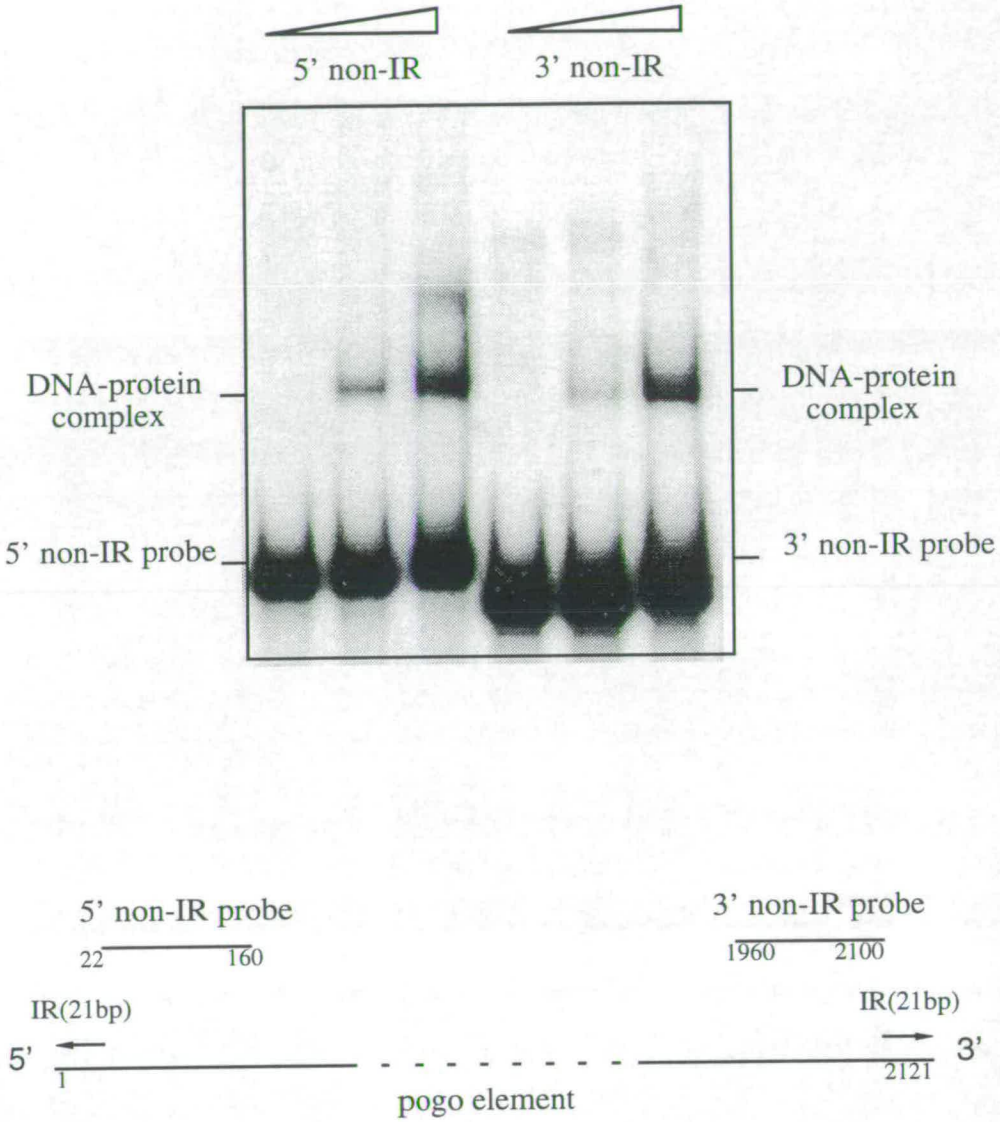


Fig. 5.6 The specific binding of pogo transposase to the 5' and 3' non-IR probes

The transposase binding activities of the 5' non-IR and 3' non-IR probes , which are the subterminal sequences from both ends of pogo without the inverted repeats, were analysed by gel retardation assay.
5' non-IR: 22-160 of pogo sequence, i.e., 5' end sequence without the inverted repeat.
3' non-IR: 1960-2100 of pogo sequence, i.e., 3' end sequence without the inverted repeat.
0, 5, and 15 µl of pogo transposase N75 (0.2 µg/µl) were used for binding to each probe.
1 ng of probe DNA was used for each reaction.

5.2.3 Computer search for transposase binding sites at both ends of *pogo* element

The DNA sequence 21-43, which was isolated from 5' end of *pogo* element and had been shown to contain the transposase binding site, was used as a quarry to search for sequences at both ends of the element, to see whether there are similar sequences at the 3' end, and whether there is more than one binding site at both ends of the element.

Fig. 5.7 shows that the computer search revealed a 14 bp and a 12 bp DNA sequences at each end of the element. Their sequences are highly conserved. L14 and L12 are identical over 12 bp sequence; L14 and R14 are identical over 14 bp except 1 bp; L12 and R12 are identical over 12 bp except 1 bp; and R14 and R12 are identical except 2 bp. So these 12-14 bp highly conserved sequences (AG)TTAG/CCTGCA/CTCG are very likely to be the transposase binding sites.

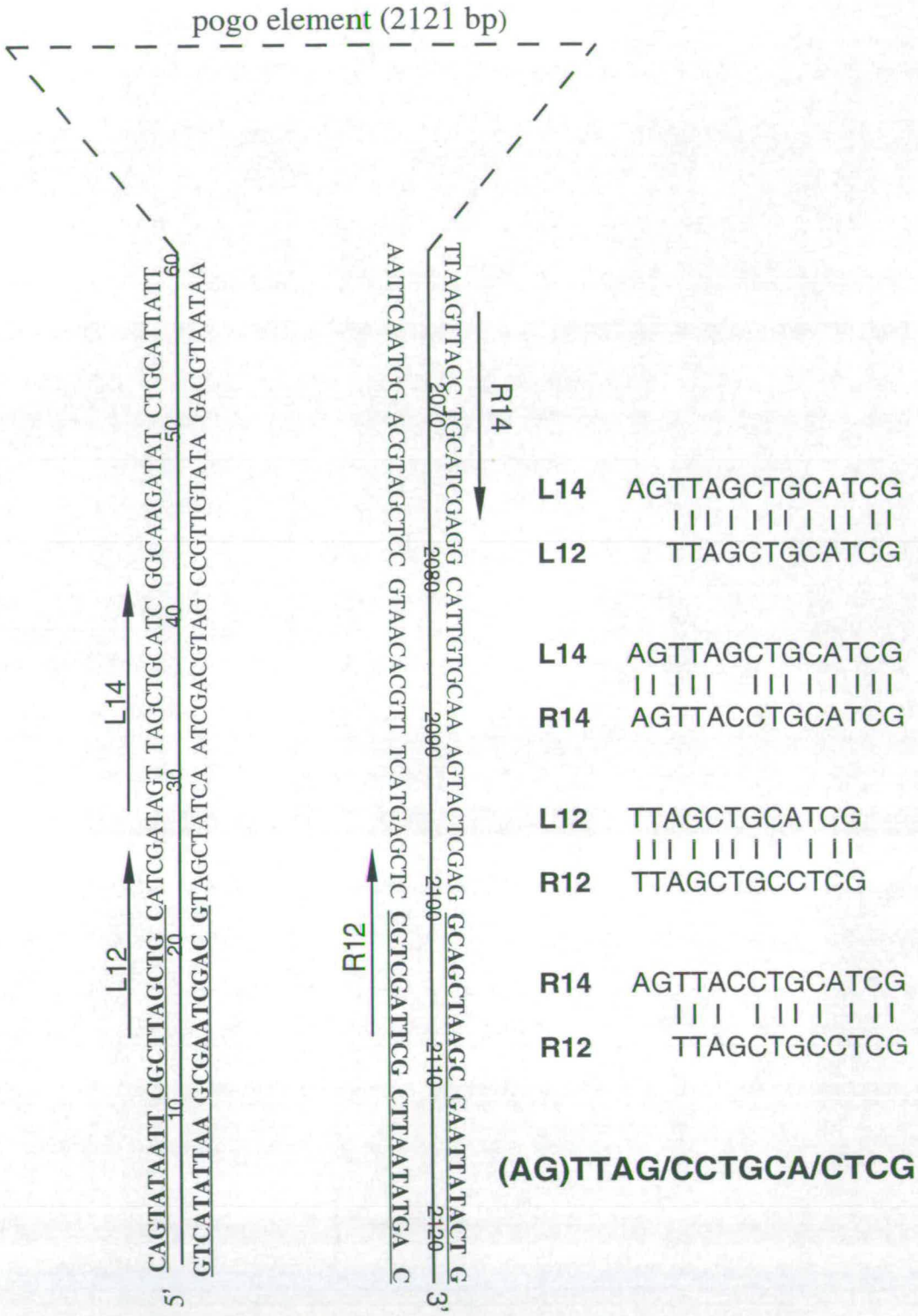
Further computer searches were conducted by using the L14 sequence as a quarry to search for the rest sequence of the element. More similar sequences were identified within sequences near the ends and also in the middle of the element.

Fig. 5.8a shows the locations of these sequences in the *pogo* element. They can be classified into four groups according to their positions: The 5' group, which includes the 14-25, 28-41 and 45-56 sequences; The 3' group, which includes the 2097-2108, 2064-2077 and 2047-2058 sequences; The 5' subterminal group, which includes the 107-119 and 193-202 sequences; and the middle position group, which includes the 1604-1614, 1302-1311 and 1155-1165 sequences.

Fig. 5.8b shows the changed bases, length of the sequences, and their percent of identity with the L14 sequence. Only those having significant identity with L14 sequence over 10 bp or longer sequences were lined out. Compared with L14 sequence, all the other sequences are subject to different terminal deletions or single nucleotide substitutions. However, only three nucleotides TGC from DNA sequence 8-10 of the L14 sequence remain unchanged. Whether these sequences are all involved in the specific binding with the transposase, and whether the three conserved nucleic acids play an crucial role in the transposase binding activity of the sequences remain to be further investigated.

5.2.4 The transposase binding ability of the isolated binding sites

To further test whether each of the these 14 bp and 12 bp binding sites itself is responsible for the specific binding to the transposase, two probes were made to include only one binding site in each. Probe 14 BS was made from a 43 bp sequence containing the 14 bp binding site flanked by polylinker DNA. Probe R12 was made by



5.7 The transposase binding sites of pogo element

DNA sequence 22-43 isolated from deletion studies was used as a quarry to search for similar sequences at both ends of the element. A 12bp and a 14 bp sequences were found at both ends . Their sequences are highly conserved, and very likely to be the transposase binding sites.

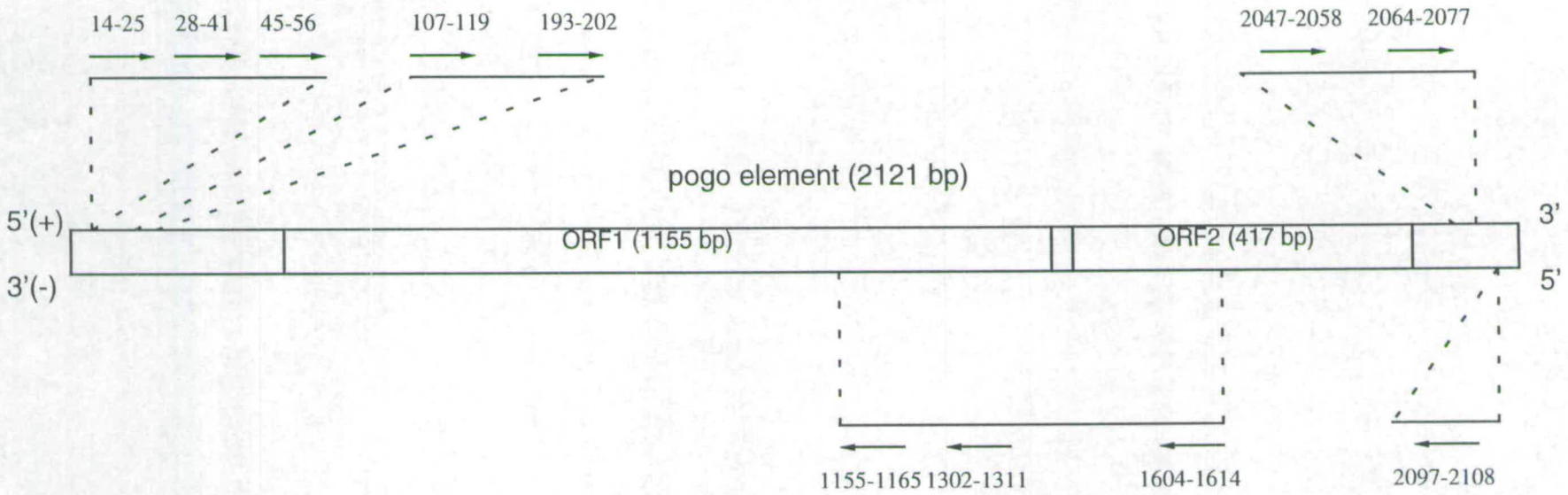


Fig. 5.8a Locations of the transposase binding sites of pogo elements
The highly conserved 14 bp sequence from the left hand side of pogo element was used as a quarry in computer search for the rest sequence of the element. Similar sequences were found at both ends and in the middle of pogo sequence. These sequences can be classified into four groups according to their positions: The 5' group, which includes the 14-25, 28-41 and 45-46 sequences; The 3' group, which includes the 2097-2108, 2064-2077 and 2047-2058 sequences; The 5' subterminal group, which includes the 107-119 and 193-202 sequences; and the middle position group, which includes the 1604-1614, 1302-1311 and 1155-1165 sequences.

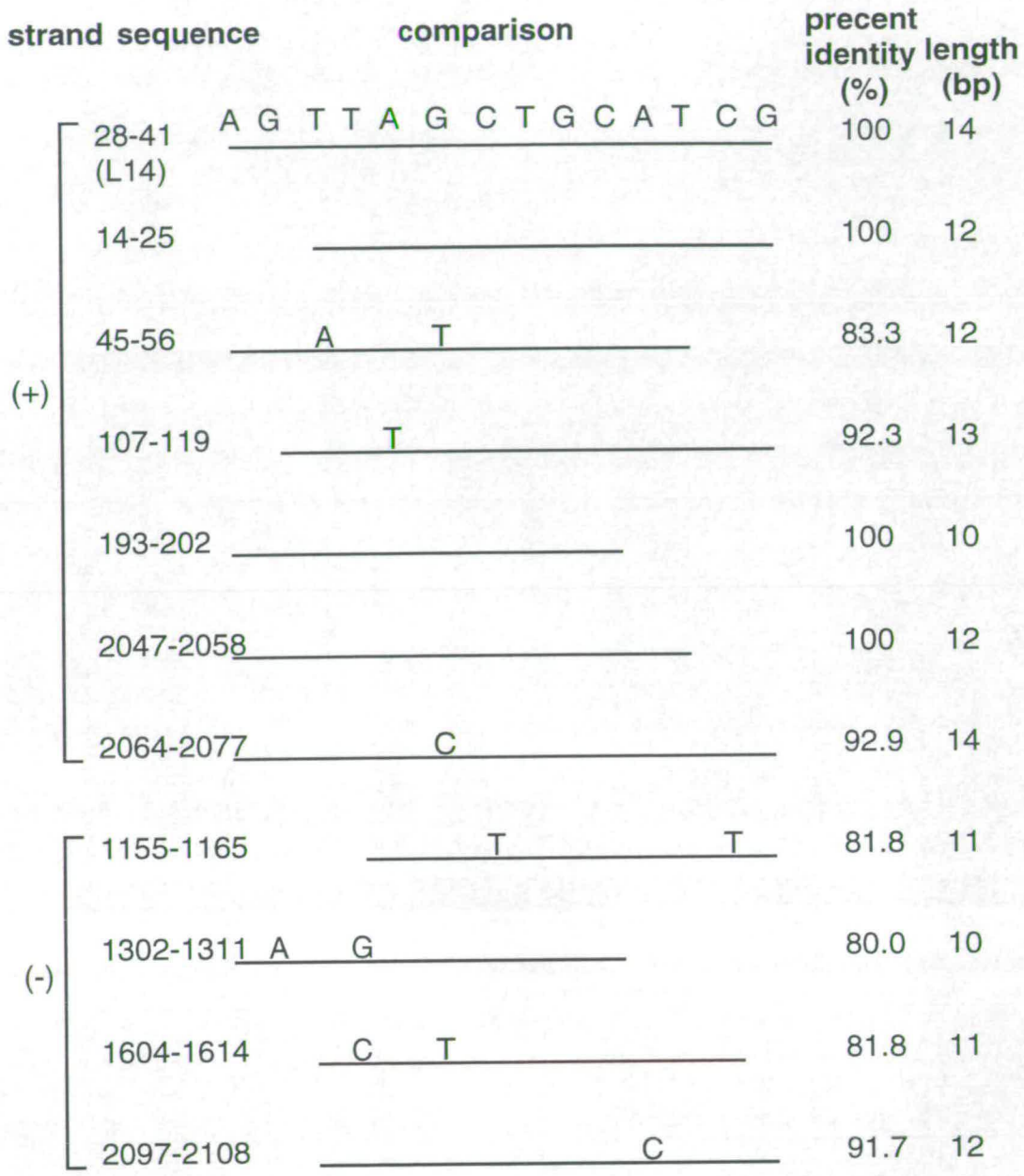


Fig. 5.8b Sequence comparison of L14 sequence with other similar sequences

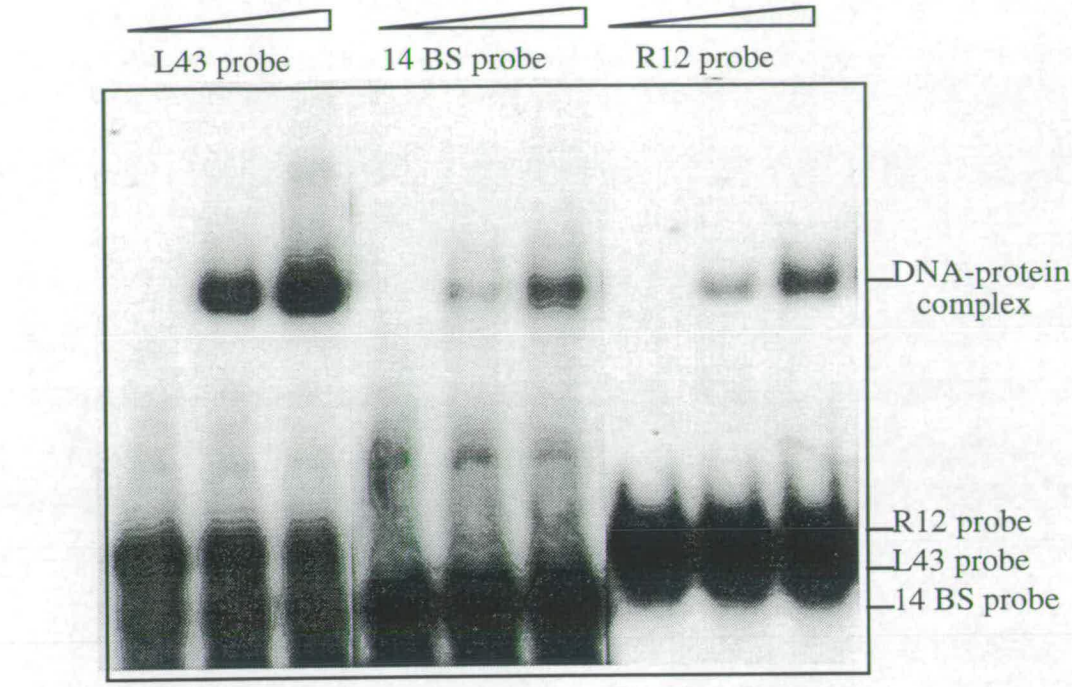
The 14 bp conserved sequence from the left hand side of pogo element was used to search the rest of the sequence. Similar sequences were found at both strands. They are located near the end and in the middle of the element. Compared to L14 sequence, each of them has different single base change and/or terminal deletions. The changed bases, length of these sequences, and their precent of indentities with that of L14 sequence are show in this diagram.

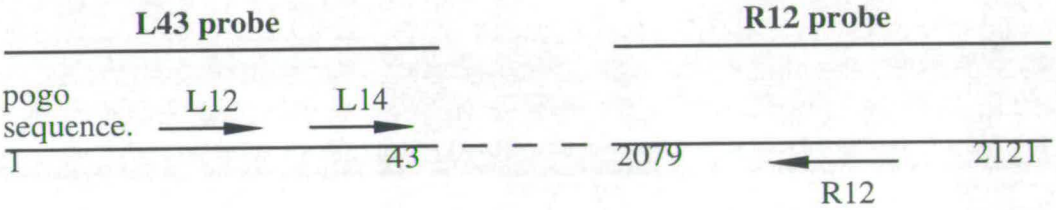
using DNA sequence from 2079 to 2121 of *pogo* element, which only contains the 12 bp binding site from 3' end of the element. The first 43 bp sequence was used as positive control, which includes both 14 bp and 12 bp binding sites from the 5' end.

Fig. 5.9 shows that the 14 BS probe and R12 probe could both bind specifically to *pogo* transposase, and their sequences are highly conserved over 12 bp. These further indicate that the 12 bp highly conserved sequences at both ends of the element are responsible for binding to the transposases, they are the transposase binding sites.

5.3 Discussion

A transposase is supposed to bind sequences specifically at both ends of the transposon, to form a highly organised synaptic transposition complex, in order to catalyse the following endonuclease cleavage and strand transfer reactions. The inverted repeats which are the same sequences located asymmetrically at both ends of the element are very likely to be the candidate for this job. The transposases of *Tc1* and *Tc3* elements from *C. elegans* have been shown to be able to bind specifically to the terminal inverted repeats of the elements. The transposase recognition site of *Tc1* is located between base pair 5-26 from the ends of the element, within the 54 bp terminal inverted repeat (Vos, et al., 1993). *Tc3* transposase binds specifically to two regions within the 462 bp terminal inverted repeat. One region is located at 13-38 bp from the end, the other is located about 180 bp from the end (Colloms, et al., 1994). However for transposons like the *P* element from *D. melanogaster* and *Ac* from *Zea mays*, this is not the case. The transposase of *P* element does not bind to the 31 bp terminal inverted repeat, instead, it binds specifically to the 10 bp consensus sequence present in inverted orientation, 16 bp from 5' inverted repeat and 4 bp from 3' inverted repeat (Kaufman et al., 1989). The transposase of *Ac* doesn't bind with the 11 bp terminal inverted repeat either. It binds specifically to the hexamer motif AAAGGG, which is repeated several times in direct or inverted orientation, in the 60 and 70 bp long segment located 100 bp inside of the 5' *Ac* terminus and 40 bp inside of the 3' terminus respectively (Kunze & Starling, 1989). For *pogo* elements, the 12 bp highly conserved sequences have been shown to be the transposase binding sites in this chapter. The outer 12 bp binding sites are located asymmetrically at the ends of the element. The first 8 bp of the sequence is within the terminal inverted repeat, while the rest 4 bp is out of it (Fig.5.7). The 21 bp inverted repeat sequence itself failed to bind with the transposase when it was tested in this chapter, because it does not contain the complete binding site. However, the 5 bp sequences immediately after the inverted repeat from both ends are identical except 1 bp. It could be the result of a point





14 BS (Binding site) probe:

GATCCGAATTCGTCGAC**AGTTAGCTGCATCG**TCTAGACTGCAGC
GCTTAAGCAGCTG**TCAATCGACGTAGC**AGATCTGACGTCGAGCT

5.9 Specific bindings of pogo transposase to DNA fragments containing different binding site(s)

14 BS and R12 probes were made to test the transposase binding activity of each of the 12 and 14 bp sequence in gel retardation assay. Probe 14 BS was made from a 43 bp sequence containing the 14 bp binding site flanked by polylinker DNA. Probe R12 was made by using DNA sequence from 2079 to 2121 of pogo element, which only contains the 12 bp binding site from 3' end of the element. A probe made from the first 43 bp sequence was used as positive control, which includes both 14 bp and 12 bp binding sites from the 5' end.

0, 5, and 10 µl of N75 protein (~0.2 µg/µl) was used for binding to each probe.

~1 ng of probe DNA was used for each reaction.

mutation occurred at one stage during evolution. In this view, the original terminal inverted repeat could be the first 26 bp sequence, instead of the present 21 bp sequence. Therefore, the transposase binding site of *pogo* element is located within this 26 bp terminal inverted repeat sequence.

The organisation of transposase binding sites differ from one transposon to another. *P* elements have one binding site at each end. *Tn7* elements have four binding sites at the left hand side and three at the right hand side. *Mu* elements have three binding sites at each end (Polard & Chandler, 1995). For the *pogo* element, several sequences closely related to the transposase binding site L14 have been found near the ends and in the middle of the element. They can be classified into four groups according to their locations (Fig.5.8a). The 5' group and the 3' group include sequences located at both ends of the element, they are more likely to be involved in forming the highly organised synaptic transposition complex with the transposase. The 5' subterminal group includes sequences located downstream of the TATA box and upstream of the coding region for the transposase (Tudor, et al., 1992). These sequences could be involved in the regulation of *pogo* transposition. The transposase, at higher concentration, could act as an repressor to bind to these sequences and block the transcription if these sequences have less affinity with the transposase than the transposase binding sites located near the ends. There is no indication as to the functions, if any, of the transposase binding sequences located in the middle of the element could be involved.

Chapter 6

THE DNA BINDING DOMAIN OF *POGO* TRANSPOSASE

6.1 Introduction

It has been shown from the works described in the previous chapters that *pogo* transposase can bind specifically to sequences at the end of the element. The 12 bp highly conserved DNA sequences at both ends of the element have been isolated and shown to be responsible for the specific interaction with the transposase. The experiments in this chapter were designed to learn which part of the transposase is involved in this sequence specific recognition and binding, which amino acids in this region are crucial for specific binding, and how much the secondary structure of the transposase contributes to this specific DNA-protein interaction.

To investigate these, two series of terminally deleted transposase derivatives, one with N-terminal deletions and the other with C-terminal deletions, have been tested for their ability to bind specifically to the transposon end sequence. Each of these terminally deleted transposases was obtained by PCR amplification of the corresponding DNA coding region, followed by cloning and protein expression in the same way as described previously for the whole transposase. The DNA binding activity of each of these proteins was analysed by gel retardation assays using the *pogo* end sequence as probe. The results show that the first 75 amino acids make up the DNA binding domain of the *pogo* transposase. Several amino acids within this DNA binding domain have been analysed by mutagenesis studies, in order to test the prediction that there is a helix-turn-helix DNA binding motif in this region (Pietrokovski & Henikoff, 1997). The results show that the positively charged basic amino acids in the recognition helix play an important role in the specific DNA-protein interaction with the transposon end sequences. So does the putative helix-turn-helix structure itself. These data therefore supports the above prediction.

6.2 Results

6.2.1 Isolating DNA binding domain of *pogo* transposase

6.2.1.1 DNA binding activity of the C-terminal regions of *pogo* transposase

A series of deletions was made by removing different amounts of the N-terminal sequence of the transposase to study the DNA binding activity of the remaining C-terminal regions of the transposase (Fig.6.1). C139 is the C-terminal 139 amino acids encoded by ORF2 of the transposon. C337 is the C-terminal 337 amino acids that include the region corresponding to the D,D35E motif which is the catalytic domain of the transposase of *Tc3* element. C396 is the C-terminal 396 amino acids that

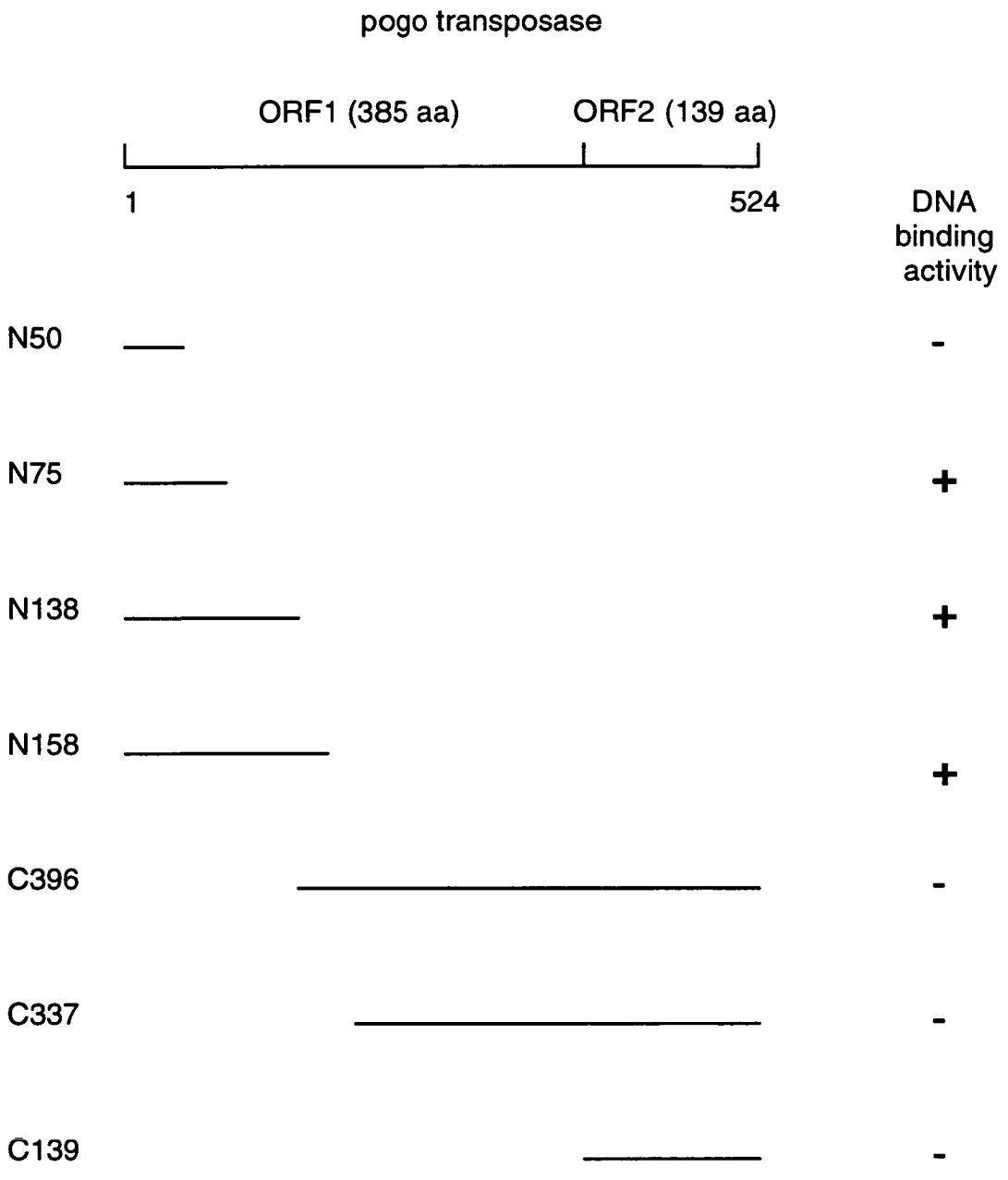


Fig. 6.1 Isolating DNA binding domain of pogo transposase by deletions
Two series of terminally deleted transposase, one with N-terminal deletions and the other with C-terminal deletions, were tested for their ability to bind specifically with the transposon end sequence. The C-terminal derivities of the transposase, C396, C337 and C193, have no specific DNA binding activity. The N-terminal derivities, N158, N138 and N75 all have specific DNA binding activity, but N50 does not. So the first 75 amino acids of the transposase make up the DNA binding domain.

corresponds to the rest of the CENP-B sequence excepting the N-terminal alpha satellite DNA binding domain (Smit & Riggs, 1996). The ability of each of these deleted proteins to *pogo* end sequence was studied by gel retardation assay.

Fig. 6.2 show the DNA binding activities of the C139, C337 and C396. Each of these proteins failed to form specific DNA-protein complex with the *pogo* end sequence. This suggests that the C-terminal region of the transposase has no specific DNA binding activity.

6.2.1.2 DNA binding activity of the N-terminal regions of *pogo* transposase

A series of N-terminal regions of *pogo* transposase were obtained by removing different amount of the C-terminus. N158 is the N-terminal 158 amino acids that include the region corresponding to the N-terminal DNA binding domain region of *Tc3* transposase. N138 is the N-terminal 138 amino acids and corresponds to the alpha satellite binding domain of CENP-B protein (Smit & Riggs, 1996). N75 and N50 are the N-terminal 75 and 50 amino acid sequences that are further deleted from the C-terminal of the N138 protein. The DNA binding activity of each of the bigger proteins was studied by gel retardation before the next deletion was made to narrow the region which is responsible for the specific binding with the *pogo* end sequence.

Fig 6.3 shows that the N-terminal sequences of the transposase can bind specifically to the transposon end when they are deleted to 158, 138, and 75 amino acids. A further deletion to 50 amino acids abolished DNA binding. This suggests that the N-terminal 75 amino acids include the DNA binding domain of *pogo* transposase.

6.2.2 Mutagenesis studies of the helix-turn-helix DNA binding motif in the DNA binding domain of *pogo* transposase

Recent work by Pietrokovski & Henikoff (1997) using a multiple-sequence alignment computer search, predicted a helix-turn-helix (HTH) DNA-binding motif in members of the *pogo* family of transposases (Fig.6.4). This predicted HTH motif for *pogo* transposase, which is from amino acid 26 to 47, is located within the N-terminal 75 amino acids DNA binding domain. To test this prediction experimentally, site-directed mutagenesis has been used to change specific amino acids within and outside the HTH motif, in order to study their role in the specific interaction of *pogo* transposase with the transposon end sequence.

To obtain each of the mutated proteins, pairs of primers were designed to change the genetic codon coding for the original amino acid to the codon for the new amino acid at both strands (Fig. 6.5). Firstly, two independent PCR reactions were

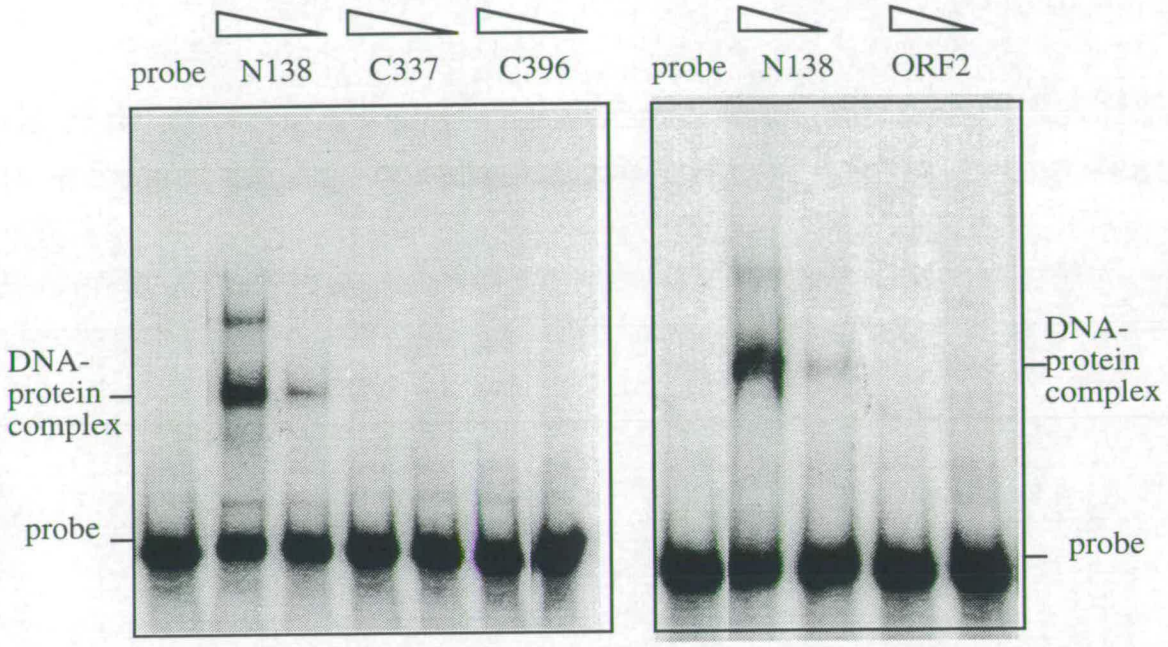


Fig. 6.2 The sequence specific DNA binding activities of the C-terminal regions of pogo transposase

A series of C-terminal regions of pogo transposase were obtained by deleting different length of the N-terminus of the transposase. The specific DNA binding activities of these deleted proteins were analysed by gel retardation assay.
probe: the first 95 bp sequence of pogo element.
N138: The first 138 amino acid sequence of pogo transposase (GST fusion protein), used as positive control.
C337: The C-terminal 337 amino acid sequence of pogo transposase (GST fusion protein).
C396: The C-terminal 396 amino acid sequence of pogo transposase (GST fusion protein).
ORF2: The C-terminal 139 amino acid sequence of pogo transposase (encoded by ORF2)(GST fusion protein).
5 and 15 µl of each of the proteins (~0.2 µg/µl) were used for each reaction.
~1 ng of probe DNA was used for each reaction.

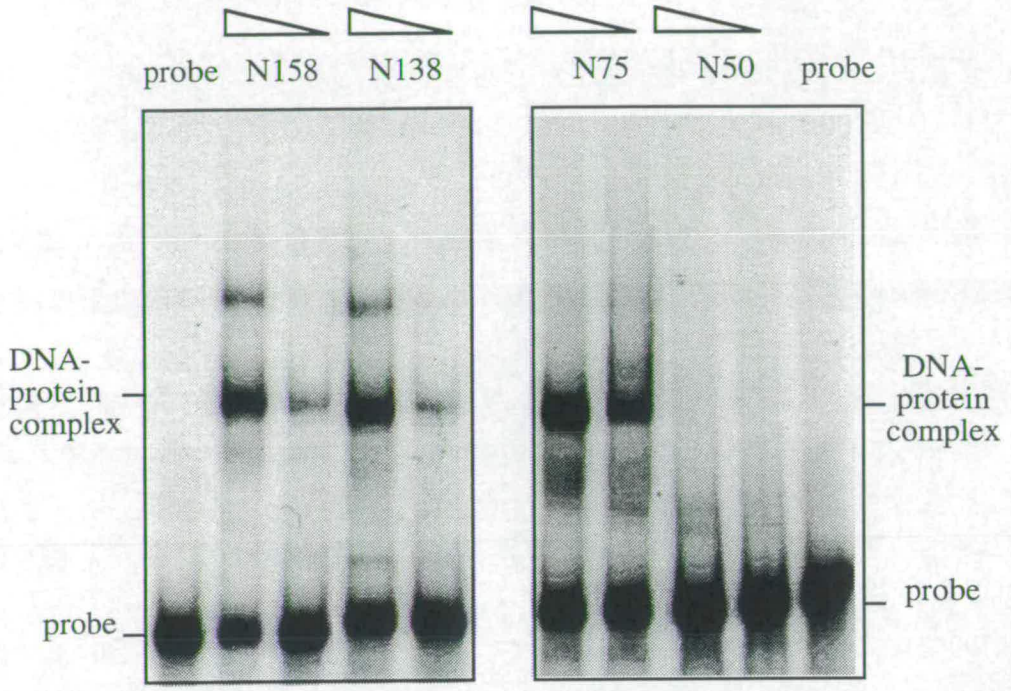


Fig. 6.3 The sequence specific DNA binding activities of the N-terminal regions of pogo transposase

A series of N-terminal regions of pogo transposase were obtained by deleting different length of the C-terminus of the transposase.The specific DNA binding activities of these deleted proteins were analysed by gel retardation assay.
probe: the first 95 bp sequence of pogo element.
N158: The first 158 amino acid sequence of pogo transposase (GST fusion protein).
N138: The first 138 amino acid sequence of pogo transposase (GST fusion protein).
N75: The first 75 amino acid sequence of pogo transposase (GST fusion protein).
N50: The first 50 amino acid sequence of pogo transposase (GST fusion protein).
5 and 15 µl of each of the N158 and N138 proteins (~0.2 µg/µl) were used for each reaction. 10 and 15 µl of each of the N75 and N50 proteins (~0.2 µg/µl)were used for each reaction.
~1 ng of probe DNA was used for each reaction.

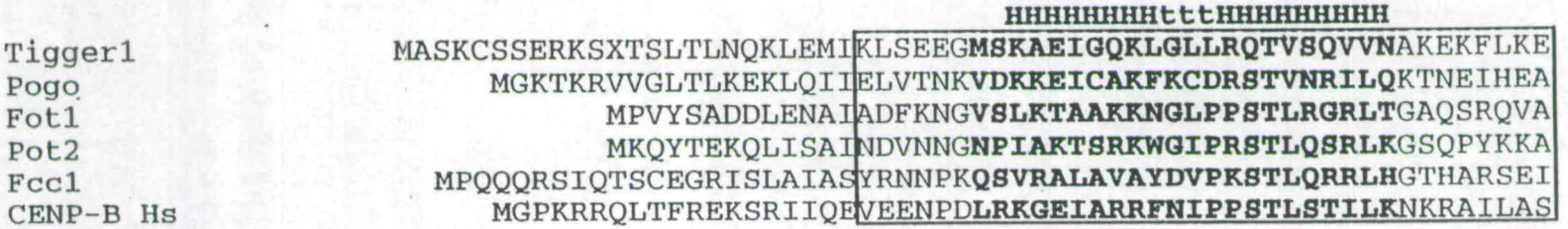


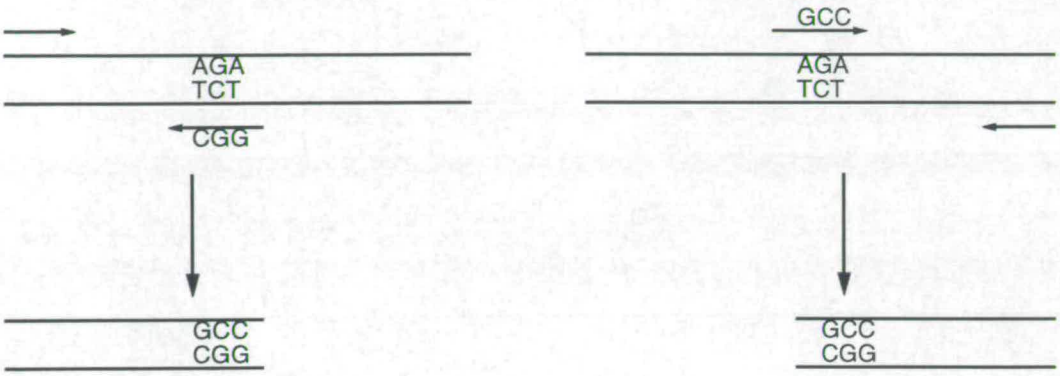
Fig. 6.4 The Helix-turn-Helix DNA binding motifs in the proteins of pogo family

(From Shmuel and Steven, 1997)

The helix-turn-helix DNA binding motifs have been predicted to be present in the N-terminal of the transposases encoded by pogo family transposons: Tigger, Pogo, Fot1, Pot2 and Fcc1. It is also present in CENP-B protein.

Amino acids making up the HTH motif are in bold letters

First PCRs:



Second PCR:

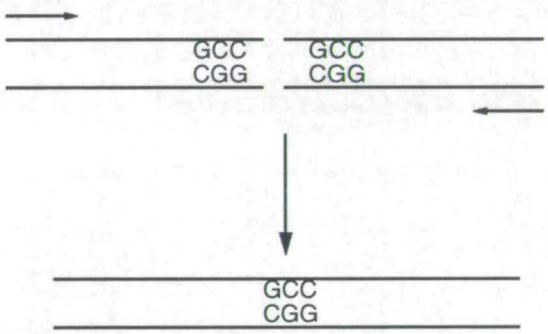


Fig. 6.5. Change genetic codons coding for different amino acids by PCR
A pair of primers were designed to change the coding sequence in both strands. Firstly, two independent PCR reactions were conducted to amplify the DNA sequences at the left and right hand sides of the changed amino acid. A second PCR was conducted by using equimolar amounts of the first two PCR products as template and the two outer primers to amplify the whole DNA sequence coding for the mutated protein. In the above example, the codon AGA coding for ariginine(39) of pogo transposase was changed to GCC for alanine.

conducted to amplify the DNA sequences at the left and right hand sides of the changed amino acid. A second PCR was conducted by using equimolar amount of the first two PCR products as template and the two outer primers to amplify the whole DNA sequence coding for the mutated protein. The final PCR product was cloned and the mutated protein was expressed in the same way as that for the wild type protein.

6.2.2.1 The role of basic amino acids in the recognition helix in the specific DNA-protein interaction

If there is a helix-turn-helix (HTH) structure in the DNA binding domain of *pogo* transposase, the amino acids in the corresponding region (amino acids 28-47) should fold as those in a general HTH structure (Fig.6.6a and b). It is the amino acids in the second helix (the recognition helix) of a HTH structure that are supposed to be involved in specific DNA-protein interactions, and positively charged amino acids in this helix are likely to be involved in this. To test this for *pogo* transposase, three basic amino acids, arginine(R)39 and arginine(R)43 which are located within the second helix, and lysine(K)48 which is one amino acid out of the second helix, were mutated separately to alanine(A). The specific DNA binding activities of these mutated proteins were analysed.

Fig. 6.7a-c shows from DNA sequence analysis that the genetic codons AGA, CGC and AAA which code for R39, R43 and K48 of *pogo* transposase had been changed to the alanine codon GCC.

Fig. 6.8 shows that the specific DNA binding activities of R39-A, R44-A and K48-A mutated proteins with the *pogo* end sequence. The R39-A and R44-A mutations knocked out the sequence specific DNA binding ability of the protein, while the K48-A mutation had much less effect. These results indicate that the positively charged basic amino acids located within the recognition helix do play an important role in the specific DNA-protein interaction.

6.2.2.2 The role of basic amino acids outside the recognition helix in the specific DNA-protein interaction

To further investigate whether positively charged amino acids anywhere else within the DNA binding domain, but outside the recognition helix, are involved in specific DNA-protein interactions between the transposase and the transposon end sequence, two amino acids, lysine(K)34 from the first helix, and arginine(R)63 from outside of the HTH motif were mutated to alanine(A). The specific DNA binding activities of the mutated proteins were analysed.

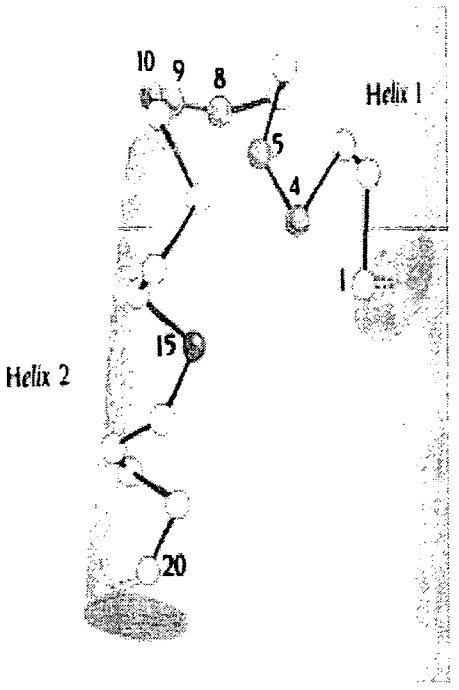


Fig 6.6a Helix-Turn-Helix motif of DNA binding proteins
(from Branden & Tooze, 1991)
A helix-turn-helix DNA binding motif is normally made up of 20 amino acids folded into two helix. The second helix (the recognition helix), is responsible for the specific binding of the protein to the DNA.

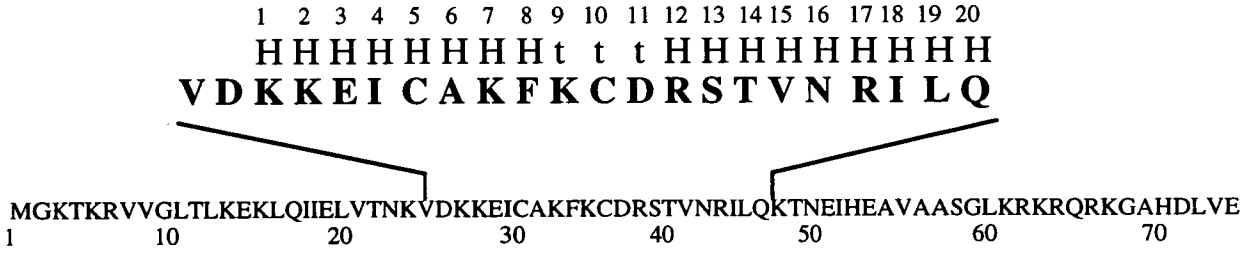


Fig.6.6b DNA binding domain and predicted helix-turn-helix motif of pogo transposase
A helix-turn-helix DNA binding motif has been predicted to be present in pogo family proteins. This predicted HTH motif for pogo transposase, which is from amino acid 26 to 47, is located within the 75 amino acids DNA binding domain isolated from deletion studies.

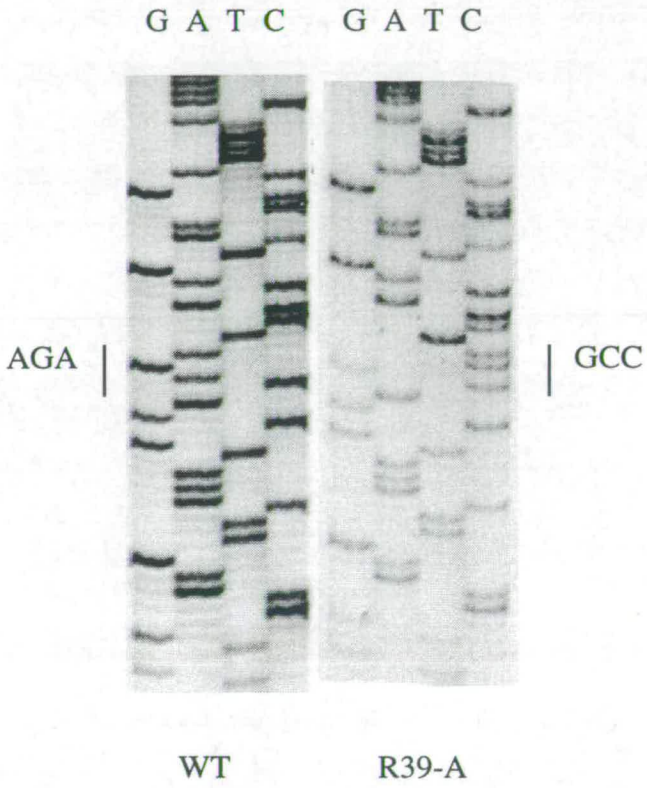


Fig.6.7a Site directed mutagenesis to change arginine(39) to alanine in the DNA binding domain of pogo transposase

The Arginine in position 39 of the pogo transposase DNA binding domain N75 was mutated to alanine by changing the codon from AGA to GCC
WT: DNA sequence coding for wild type N75 protein
R39-A: DNA sequence coding for the arginine(39) to alanine mutated N75 protein

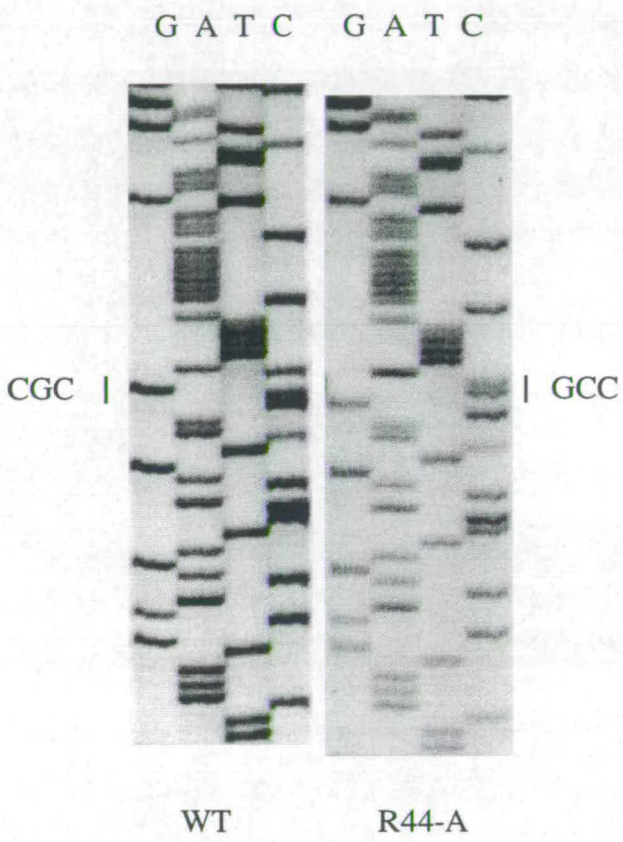


Fig. 6.7b Site directed mutagenesis to change arginine(44) to alanine in the DNA binding domain of pogo transposase

The arginine in position 44 of the pogo transposase DNA binding domain N75 was mutated to alanine by changing the codon from CGC to GCC
WT: DNA sequence coding for wild type N75 protein
R44-A: DNA sequence coding for the arginine(44) to alanine mutated N75 protein

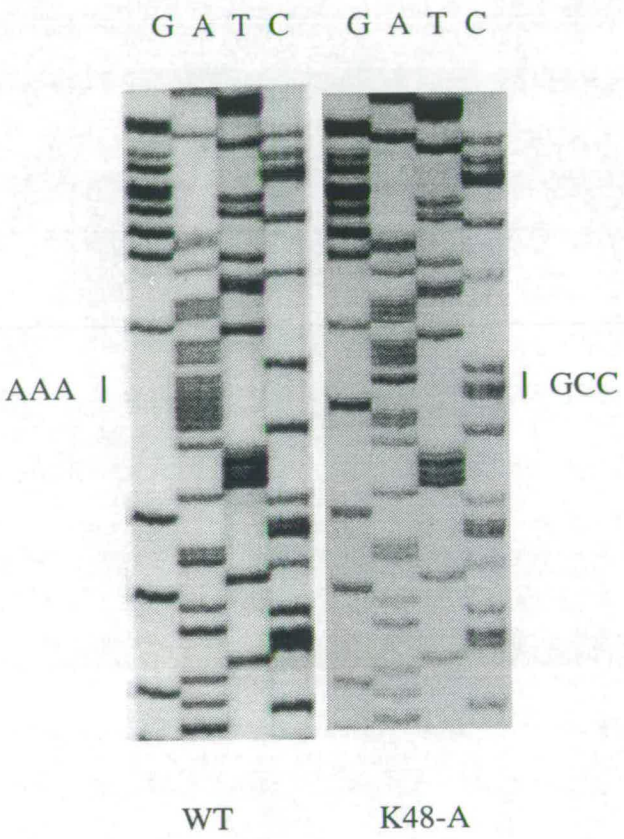


Fig. 6.7c Site directed mutagenesis to change lysine(48) to alanine in the DNA binding domain of pogo transposase

The lysine in position 48 of the pogo transposase DNA binding domain N75 was mutated to alanine by changing the codon from AAA to GCC
WT: DNA sequence coding for wild type N75 protein
R48-A: DNA sequence coding for the lysine(48) to alanine mutated N75 protein

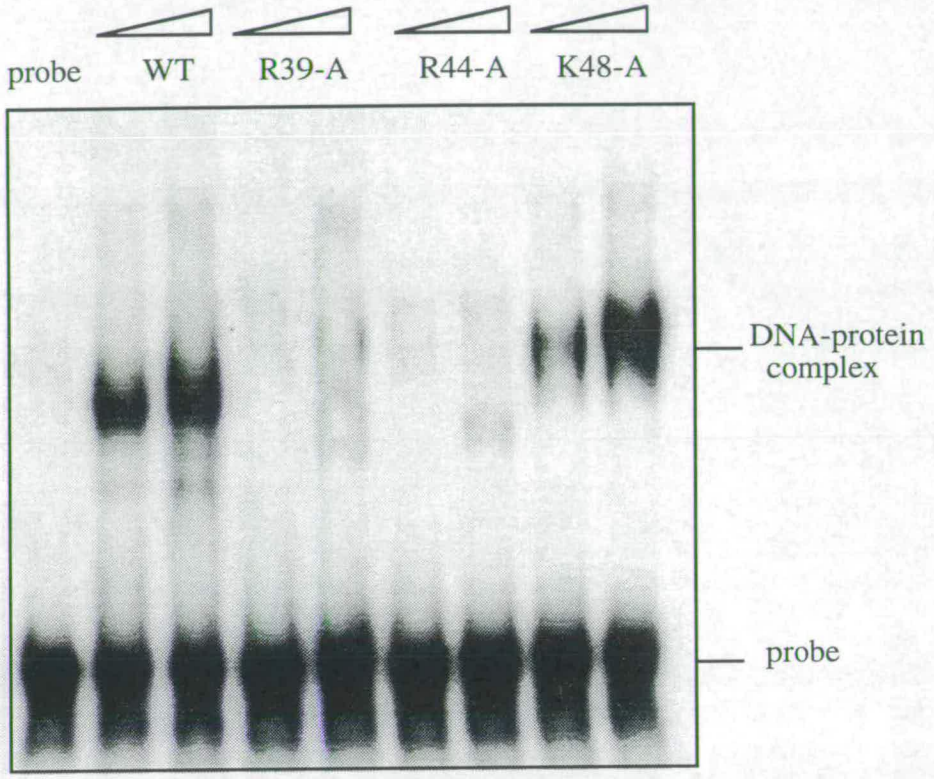


Fig. 6.8 The DNA binding activities of R39-A, R44-A and K48-A mutated N75 proteins

The amino acids R39 and R44, which are located within the predicted recognition helix, and K48 which are located just one amino acid out of the recognition helix, were mutated to alanine(A). The sequence specific DNA binding activities of these mutated proteins were analysed by gel retardation assay.

Probe: 1-95 of pogo sequence.
WT: wild type N75 DNA binding domain of pogo transposase.
R39-A: arginine(39) to alanine mutated N75 protein.
R44-A: arginine(44) to alanine mutated N75 protein.
K48-A: lysine(48) to alanine mutated N75 protein.

5 and 10 µl of each protein (~0.2 µg/µl) were used for binding with the probe.
~1 ng of probe DNA was used for each reaction.

Fig. 6.9a and b show from DNA sequencing analysis that the genetic codons AAG and AGA which coding for K34 and R63 of *pogo* transposase were changed to the alanine codon GCC.

Fig. 6.10 shows that the specific DNA binding activities of the K34-A and R63-A mutated proteins were unaffected by these mutations, as the mutated proteins could bind as well to the *pogo* end sequence as did the wild type N75 protein. These results indicate that positively charged amino acids located outside the recognition helix, even within the first helix, are not required for the specific interaction with the transposon end sequence.

6.2.2.3 The role of the helix-turn-helix structure in the specific DNA-protein interaction

To investigate how much the helix-turn-helix structure itself contributes to the specific DNA-protein interaction of the transposase with the transposon end sequence, two mutations were made to change cysteine(C)32 in the first helix and valine(V)42 in the second helix to proline. Since proline is unable to form a hydrogen bond from its main-chain nitrogen, the introduction of proline into HTH motif would be expected to disrupt the structure of the protein. The DNA binding activities of these mutated proteins were analysed by gel retardation assay.

Fig. 6.11a and b show from DNA sequencing analysis that the genetic codons TGT and GTC which code for C32 and V42 of *pogo* transposase were changed to the proline codon CCC.

Fig. 6.12 shows that the specific DNA binding activities of C32P and V42P mutated proteins with *pogo* end sequence were greatly reduced compared to those of the wild type protein. These mutated proteins had much less affinity for the DNA probe and that the DNA-protein complexes moved much more slowly probably because of misfolding of the mutated proteins.

6.3 Discussion

In a transposition process, the transposase is responsible for bringing together the two transposon ends and the target DNA, then catalysing the endonuclease cleavage and strand transfer reactions. These different jobs have been shown to be carried out by different domains of the transposase. The N-terminal regions have been shown to be the DNA binding domains for almost all transposase studied so far, while the central regions which containing the highly conserved D,D(35)E motif are the catalytic domains essential for the endonuclease cleavage and strand transfer reactions. The C-terminal regions of some transposases have been shown to contain dimerisation domains that are responsible for the dimerisation of the transposase monomers. The

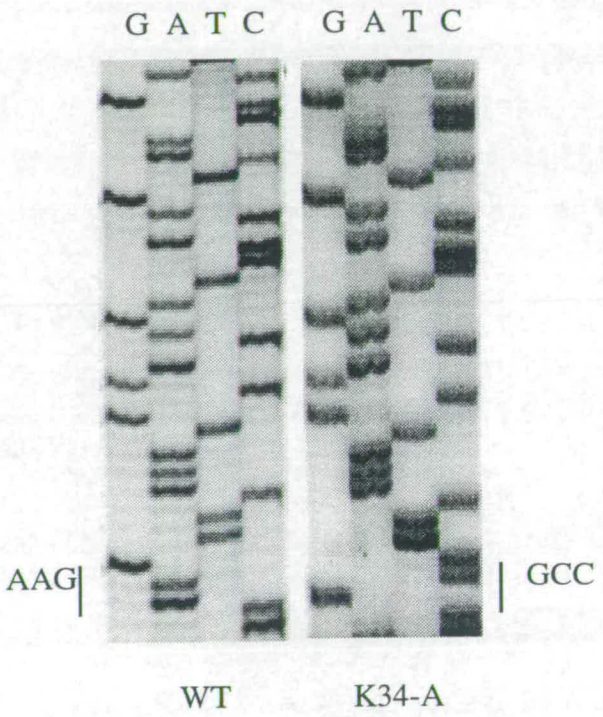


Fig. 6.9a Site directed mutagenesis to change lysine (34) to alanine in the DNA binding domain of pogo transposase

The Lysine in position 34 of the pogo transposase DNA binding domain N75 was mutated to alanine by changing the codon from AAG to GCC
WT: DNA sequence coding for wild type N75 protein
K34-A: DNA sequence coding for the lysine(34) to alanine mutated N75 protein

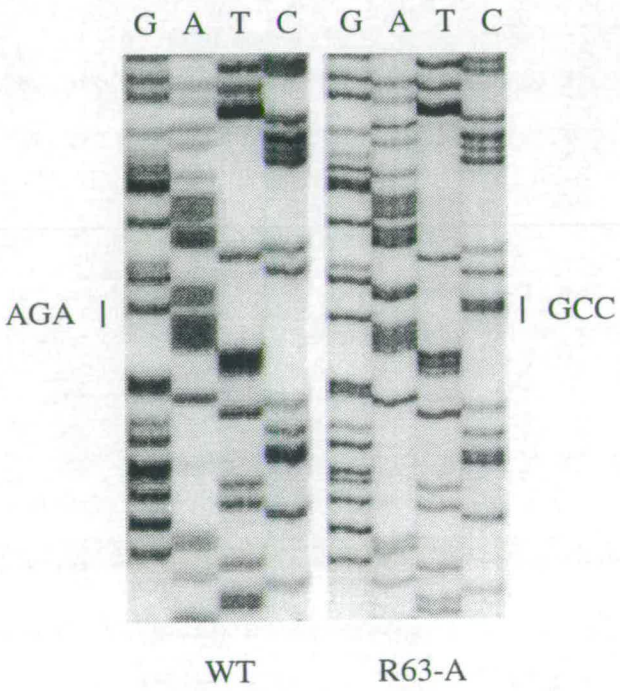


Fig. 6.9b Site directed mutagenesis to change arginine(63) to alanine in the DNA binding domain of pogo transposase

The arginine in position 63 of the pogo transposase DNA binding domain N75 was mutated to alanine by changing the codon from AGA to GCC
WT: DNA sequence coding for wild type N75 protein
R63-A: DNA sequence coding for the arginine(63) to alanine mutated N75 protein

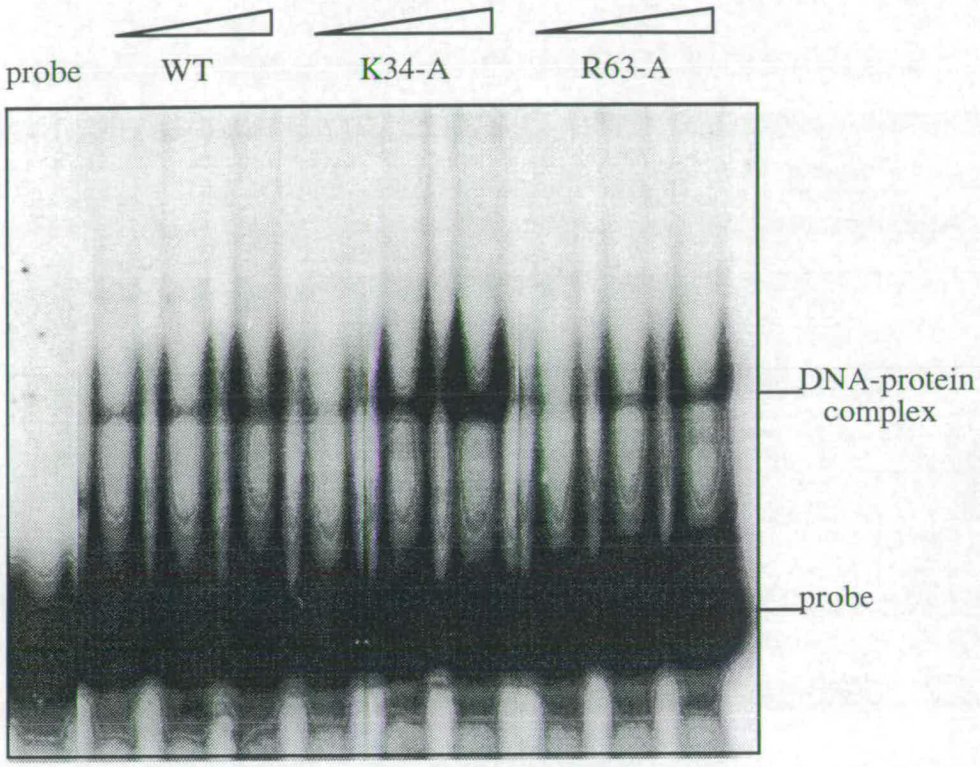


Fig. 6.10 The DNA binding activities of K34-A and R63-A mutated N75 proteins

The amino acids K34 from the first helix and R63 from outside the HTH motif were mutated to alanine(A). The DNA binding activities of the mutated proteins were analysed by gel retardation assay.
Probe: 1-95 of pogo sequence.
WT: wild type N75 DNA binding domain of pogo transposase.
K34-A: lysine(34) to alanine mutated N75 protein.
R63-A: arginine(63) to alanine mutated N75 protein.
1, 5 and 10 µl of each protein(~0.2 µg/µl) were used for binding with the probe.
~1 ng of prob DNA was used for each reaction.

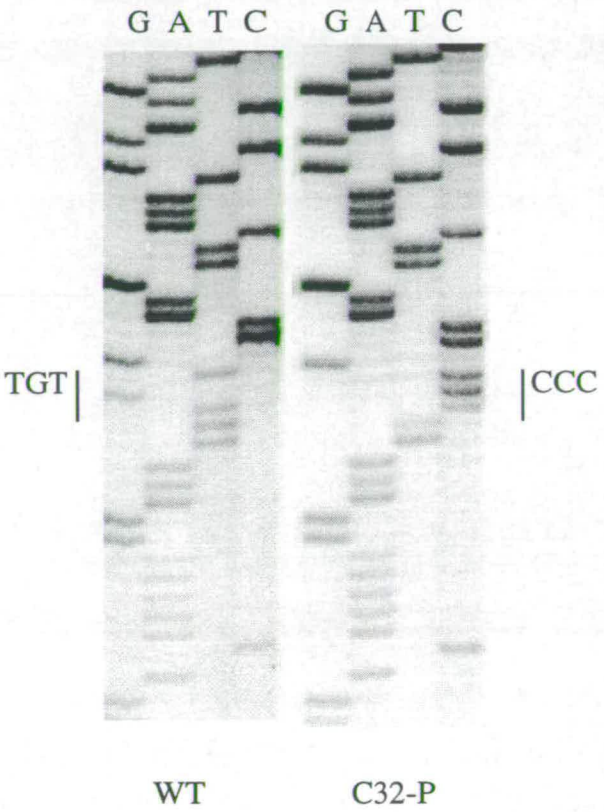


Fig 6.11a Site directed mutagenesis to change cysteine(32) to proline in the DNA binding domain of pogo transposase

The cysteine in position 39 of pogo transposase DNA binding domain N75 was mutated to proline by changing the codon from TGT to CCC
WT: DNA sequence coding for wild type N75 protein
C32-P: DNA sequence coding for the cysteine(32) to proline mutated N75 protein

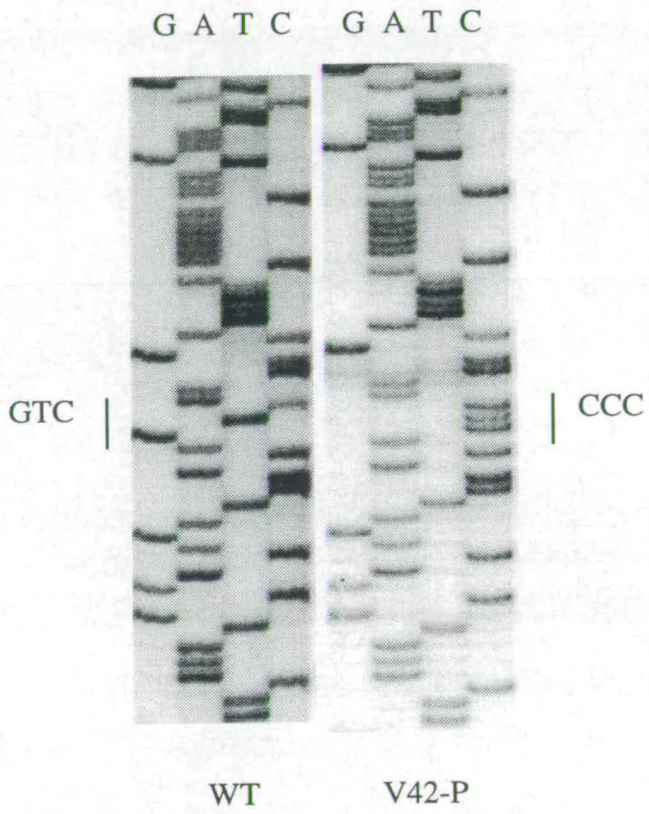


Fig. 6.11b Site directed mutagenesis to change valine(42) to proline in the DNA binding domain of pogo transposase

The valine in position 42 of pogo transposase DNA binding domain N75 was mutated to proline by changing the codon from GTC to CCC
WT: DNA sequence coding for wild type N75 protein
V42-P: DNA sequence coding for the valine(42) to proline mutated N75 protein

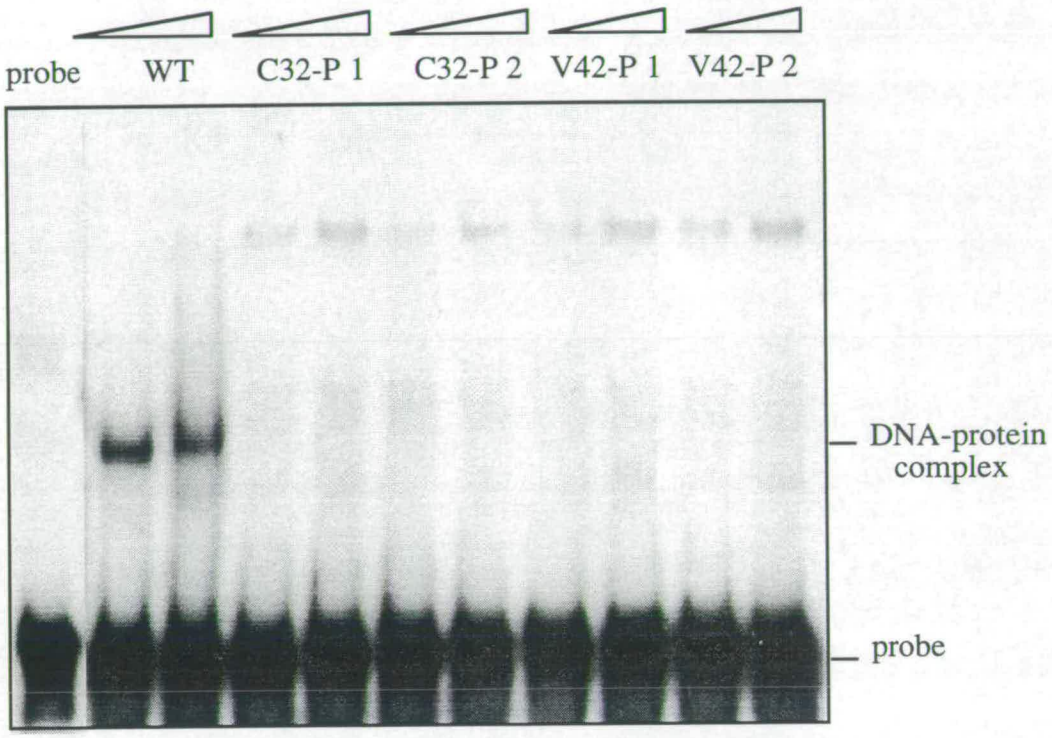


Fig. 6.12 The DNA binding activities of C32-P and V42-P mutated N75 proteins

The amino acids C32 in the first helix and V42 in the second helix were mutated to proline. The DNA binding activities of these mutated proteins were analysed by gel retardation assay.
probes: 1-95 of pogo sequence.
WT: wild type N75 DNA binding domain of pogo transposase.
C32-P 1&2: cysteine(32) to proline mutated N75 protein.
V42-P 1&2: valine(42) to proline mutated N75 protein.
5 and 10 µl of each protein (~ 0.2 µg/µl) were used for binding with the probe.
~1 ng of probe DNA was used for each reaction.

results reported here show that for the *pogo* element, the DNA binding domain is present in the first 75 amino acids of the transposase, and that the C-terminal region has no specific DNA binding activity. The central region of the transposase has an amino acid sequence similar to that of the D,D(35)E motif (Smit & Riggs, 1996), and might be responsible for catalysing the transposition reaction. There is no indication as to whether the transposase contains a dimerisation domain. However, there is more than one binding site at each end of the element. As the transposase can bind to these sequences specifically, it is not difficult to imagine that the easiest way for the transposase to bring together the two ends of the element, is to form dimers or multimers, in order to form the highly organised synaptic DNA-protein complex. The location of the dimerisation domain, and how transposase monomers interact with each other are still questions that remain to be investigated further.

Several classes of double-stranded DNA binding proteins have been identified, including helix-turn-helix motifs, helix-loop-helix motifs, zinc fingers, leucine zippers and so on. The HTH is the best characterised class among of these. Computer searches also revealed the HTH blocks in the transposase sequences of *Tc1-mariner* family (Pietrokovski & Henikoff, 1997). This has been supported by the experimental evidences that these blocks are located in the DNA binding domains of *Tc1* and *Tc3* transposases (Colloms et al., 1994; Vos and Plasterk, 1994). For the *pogo* element, the predicted HTH motif (Pietrokovski & Henikoff, 1997) is just located in the 75 amino acid DNA binding domain identified by the deletion studies in the thesis. Mutagenesis studies show that only the positively charged amino acids located within the recognition helix play an important role in the specific DNA-protein interaction with the transposon end sequence, not the same residue located elsewhere in the sequence. These data therefore supported the above prediction. The HTH structure itself has also been shown to play a very important role in the specific DNA-protein interaction. When the HTH structure was disrupted by the proline substitution, the protein had much less affinity for the transposon end sequence. This is because the HTH structure provides the best position for the amino acid side chains to contact the bases of the DNA major groove specifically. When the structure of the protein was disrupted, it's much more difficult for the protein to reach the DNA properly, so their affinity was greatly reduced.

Chapter 7

THE ENDONUCLEASE ACTIVITY OF *POGO* TRANSPOSASE

7.1 Introduction

A transposase is supposed to play a major part in catalysing transposition reactions including endonuclease cleavage and strand transfer. After recognition and binding sequence-specifically to the ends of the transposon, the transposase cleaves the transposon ends from the donor DNA. For transposons which transpose via a cut and paste mechanism, such as *Tc1* and *mariner*, the cleavage occurs at both ends of the elements, resulting in excised elements which can then insert into the target sites in the subsequent strand transfer reaction (Vos et al., 1996; Smith, et al., 1997). This endonuclease cleavage activity of a transposase can be tested by an *in vitro* excision assay. In this assay, a supercoiled plasmid DNA containing the transposon is exposed to the transposase in the presence of the divalent cations Mg^{2+} or Mn^{2+}. After the reaction, the samples are run in an agarose gel and visualised by ethidium bromide staining and southern blotting. Since *pogo* is structurally similar to *Tc1* and *mariner* elements, and also is supposed to transpose via a cut and paste mechanism, the endonuclease activity of *pogo* transposase was tested in this chapter.

7.2 Results

7.2.1 Making the pUC-pogo construct

Since a supercoiled plasmid containing the transposon is required as the substrate for the transposase in an excision assay and the available plasmid pogoR11XC only contains the *pogo* without the terminal inverted repeats (Tudor et al., 1992), a pair of primers which include the terminal inverted repeats and the TA dinucleotide target site sequence were designed and used to amplify the *pogo* element from pogoR11XC by PCR. This PCR product was cloned into a pUC18 vector via the *BamHI* and *XbaI* sites to make the pUC-pogo construct.

7.2.2 Determining the endonuclease activity of *pogo* transposase by the excision assay

The endonuclease activity of *pogo* transposase was determined by incubating the pUC-pogo DNA and the *pogo* transposase together in the presence of Mn^{2+} as described in 2.2.3.5. After the reaction, the samples were run in an agarose gel and visualised by ethidium bromide staining and southern blotting using *pogo* DNA as the probe.

Fig. 7.1a and b show, from the photograph of the agarose gel and the autoradiograph of the southern blotting, that the amount of 4.8 kb DNA which corresponds to the linearised pUC-pogo increases as the incubation time increases.

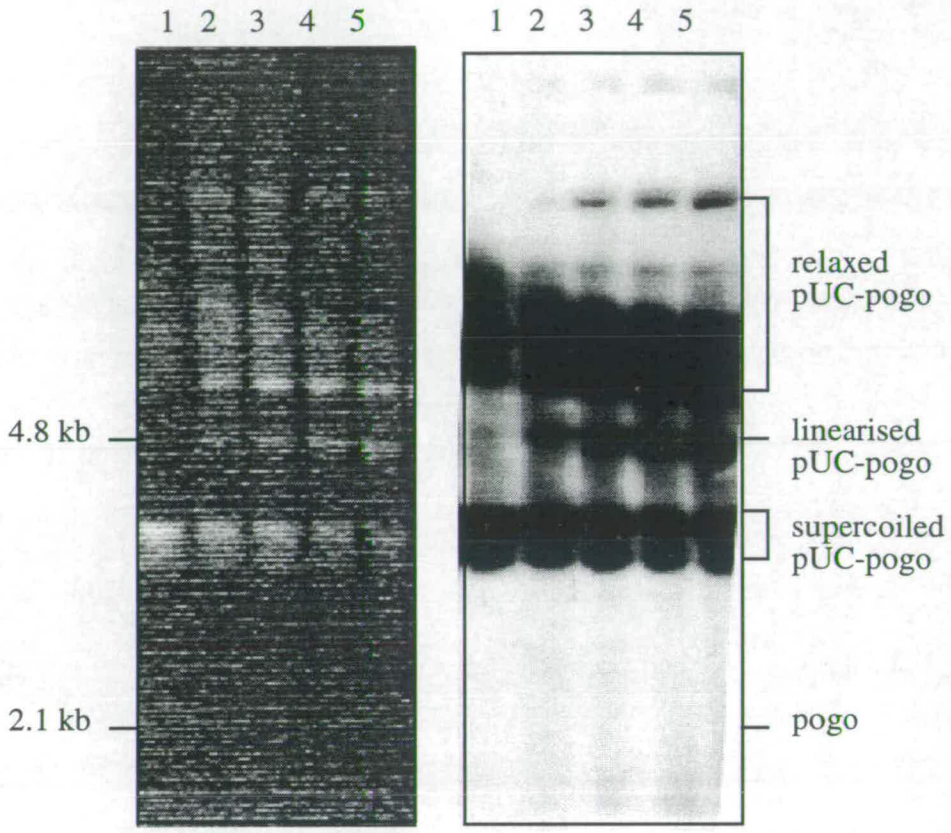


Fig. 7.1 Determining endonuclease activity of pogo transposase by excision assay

Approximately 200 ng of pUC-pogo DNA was incubated with 100 ng of pogo transposase in the presence of Mn^{2+} at 37 °C for 15, 30, 45, and 60 min respectively. The samples were run in 1% agarose gel and detected by ethidium bromide staining and southern blotting using pogo DNA as probe.

1. pUC-pogo DNA was incubated in the reaction buffer at 37 °C in the absence of the transposase for 60 min.

2-5. pUC-pogo DNA was incubated in the reaction buffer at 37 °C in the presence of pogo transposase for 15, 30, 45, and 60 min respectively.

However, no DNA band representing the 2.1 kb free *pogo* element could be seen even after southern blotting, indicating that no *pogo* elements had been excised out.

7.3 Discussion

To excise an element from its donor DNA, the transposase has to catalyse both 3' and 5' end DNA cleavage reactions. The result of the excision assay of *pogo* element showed that the amount of linearised pUC-pogo increased as the reaction time increased. The linearisition of the plasmid could due to one of the two reasons: firstly, it might due to the endonuclease activity of *pogo* transposase which somehow only cut the 5' or 3' end of the element; secondly, it might due to a contaminating endonuclease activity which cut the plasmid at non-specific sites. The linearised plasmid could be further analysed by digesting it with two enzymes which cut at different ends of the element in independent reactions. If the first is true, free elements should be released by either one or both of these digestions, and the end which was cut by the transposase could therefore be determined. If the second is true, a smear representing DNA of different sizes should be seen. The 4.8 kb linearised pUC-pogo plasmid was digested with *BamHI* and *XbaI*, which are the cloning sites of the *pogo*, respectively, then analysed by southern blotting, but no free elements could be seen in either reaction (data not shown). So it is more likely that pUC-pogo was linearised by a contaminating nonspecific endonuclease activity.

A highly organised synaptic complex including the two transposon ends and the transposase is supposed to be required for the transposase to catalyse the transposition reaction, the flanking DNA of the transposon might also contribute to the assembly of the complex. Since the pUC-pogo construct only has the TA dinucleotide in addition to the *pogo* element, another construct was made by PCR amplifiying the *pogo* element together with the 25 bp flanking DNA from *pogo* λR11 (Tudor, et al., 1992) and cloning the PCR product into a pGEM vector. This construct was also used as substrate to test the endonuclease activity of *pogo* transposase in the excision assay, but no excised transposon was detected (data not shown). This result suggests that either more flanking DNA is required for the proper assembly of the synaptic complex to enable *pogo* transposase to catalyse the DNA cleavage reaction, or the transposase itself was not active in this assay.

Chapter 8

DISCUSSION

8.1 General discussion

8.1.1 *pogo* transposition

Transposable element *pogo* from *Drosophila melanogaster* is structurally similar to the other DNA mediated elements such as *P, hobo, Tc1, Tc3,* and *Mariner* elements in having short terminal inverted repeats, creating target site duplication upon insertion, and having open reading frames coding for the transposase. So *pogo* is considered to be a member of the same class as the above elements, and to transpose from DNA to DNA probably also *via* a cut and paste mechanism (Tudor, et al., 1992).

A transposase plays a key role in the transposition process by catalysing the transposition reactions. cDNA analysis had shown that the two open reading frames of *pogo* code for a single polypeptide which could be the putative transposase (Tudor, et al., 1992). However no more information about this transposase was available before the start of this thesis. The experiments in this thesis were designed firstly to obtain *pogo* transposase by cloning the ORFs into a pGEX vector, and express the transposase as a GST fusion protein in *E.coli.* Since sequence specific DNA binding activity is a character of all transposases studied so far, so the specific DNA binding activity of the expressed *pogo* transposase was tested by gel retardation assays in the presence of specific and non-specific competitors. *pogo* transposase was shown to be able to bind specifically with the transposon end sequences. Then, the transposase binding site of the element and the DNA binding domain of the transposase, both are directly involved in the specific DNA-protein interaction, were identified. The transposase binding sites tested were shown to be a 12 bp consensus sequence, there are also several copies of the 12 bp consensus sequence located near the ends, in the 5' subterminal region, and in the middle of the element. The DNA binding domain was shown to be located in the N-terminal 75 amino acid region of the transposase. These data provide useful information for further studies of *pogo* transposition, such as how the active DNA-protein complex is assembled as an initial step of transposition; whether *pogo* transposase is the only protein required for catalysing the transposition reaction; how *pogo* transposition is regulated; and so on. These data can also be used to compare with the corresponding data from other elements, as this will help us to have a better understanding of the general transposition mechanism.

8.1.1.1 Transposase binding sites and DNA binding domain

The transposase binding site of the transposon and the DNA binding domain of the transposase are directly involved in the DNA-protein complex. So their sequences

and/or structures will have important effects on the specific DNA-protein interaction of the transposase with the transposon end sequences.

The transposase binding sites of the *pogo* element have been shown to be a 12 bp consensus sequence by testing the transposase binding ability of deleted end sequences of the element. Computer search show that there are several copies of the 12 bp sequence at different places of the element, and they can be divided into four groups according to their locations: the 5' group, the 5' subterminal group, the middle position group, and the 3' group. The 5' group and the 3' group, each contains three copies of the 12 bp consensus sequence, are located in a region of 14-56 bp from the 5' end and in a region of 14-74 bp from the 3' end respectively. These six binding sites are very likely to be involved in the synaptic DNA-protein complex in the presence of the transposase (see later). The 5' subterminal group, which contains two copies of the 12 bp consensus sequence, is located down stream of the TATA box and upstream of the coding region for the transposase (Tudor, 1992), and might be involved in the regulation of *pogo* transposition (see later). The function of the transposase binding sites located in the middle of the transposon is difficult to predict.

The specific DNA binding domain of *pogo* transposase has been shown to be located within the first 75 amino acid region of the transposase by deletion studies in this thesis. A helix-turn-helix DNA binding motif has recently been predicted to be located just within the DNA binding domain in between amino acid 26 and 47 (Pietrokovaski & Henikoff, 1997). The N-terminal 50 amino acid region of *pogo* transposase which includes the entire HTH sequence and three extra amino acids at its C-terminus was also tested for its DNA binding activity, but failed to bind specifically to the transposon end sequence. This might be because the three extra amino acids at the C-terminus of the HTH sequence are not sufficient to keep the HTH motif folded properly, it is possible that more than the HTH is required for the binding. The HTH DNA binding motif has also been predicted to be presence in the DNA binding domain of *IS30, Tc1* and *Tc3* transposases (Pietrokovski, 1996; Stalder et al., 1990; Pietrikovski & Henikoff, 1997), however no direct evidence has ever shown that the HTH motif is the motif responsible for the specific DNA binding ability of the DNA binding domain. In this thesis, mutagenesis studies show that when a single basic amino acid in the recognition helix was substituted, the DNA binding domain lost its DNA binding activity; and the same result was obtained when a single amino acid within the HTH was substituted by proline which should disrupt the HTH structure. These data clearly indicate that the HTH motif within the DNA binding domain is responsible for the specific DNA binding activity of the transposase.

8.1.1.2 Specific DNA-protein interaction of *pogo* transposase with *pogo* end sequences

Sequence specific DNA binding activity has been shown to be an essential character of all transposases studied so far. The high affinity of the transposase for the transposon end sequences enables the transposase to recognise and interact specifically with the two transposon ends when the transposon is still embedded in the genomic DNA. *pogo* transposase has been shown to be able to bind to the transposon end sequence in the presence of 100x molar excess of nonspecific competitor, indicating it has a much higher affinity for the element end sequence than for other DNA sequences, so the binding is sequence specific.

8.1.1.3 The architecture of *pogo* DNA-protein complex

As an initial step of transposition, a transposase has to recognise and bind specifically to the ends of its transposon to form a synaptic DNA-protein complex (Haniford, et al., 1991; Haniford & Kleckner, 1994; Sakai et al., 1995; Mizuuchi, et al., 1992; Bainton et al., 1991, 1993; Sarnovsky, et al., 1996). This complex is supposed to be essential for catalysing the transposition reaction, since only in this complex can the transposase become activated. The composition and architecture of the complex are different from element to element, as is shown by *Tn10*, *Mu* and *Tn7* DNA-protein complex (Balland & Nancy, 1996; Mizuuchi, et al., 1992; Rogers et al., 1986; Waddell & Craig, 1988; Kubo & Craig, 1990). This depends on the number of transposase binding sites and the number of transposase monomers involved. Some host protein, such as IHF protein that forms part of the *Mu* and *Tn10* complexes, and/or some DNA sequences from the element, such as the IAS sequences that are part of the *Mu* complex, might also be involved (Signon & Kleckner, 1995; Mizuuchi, et al., 1992; Mizuuchi & Mizuuchi, 1989;, Leung et al., 1989; Surette et al., 1989).

For *pogo*, each of the 5' end and the 3' end transposase binding site groups contains three transposase binding sites located near the end of the element. The organisation of these six transposase binding sites makes it very like that of phage *Mu*. At the ends of *Mu*, there are also six transposase binding sites, three at each end, named L1, L2, L3 on the left and R1, R2, R3 on the right from the end. *Mu* transposase monomers can bind specifically with each of these binding sites, while in the synaptic DNA-protein complex, the transposase forms a tetramer which only stably binds to the three endmost sites, L1, R1 and R2. Two monomers of the tetramer catalyse the DNA cleavage reaction, while the other two catalyse the strand transfer reaction (Allison & Chaconas, 1992; Mizuuchi et al., 1992). For *pogo*, it is also possible for the transposase to bind to all the six binding sites at one stage in

transposition, and bind to only some of them at another stage. The interactions between transposase monomers might also be changeable. However, more information about the protein-protein interaction of *pogo* monomers and the relative affinity of each transposase binding site for the transposase have to be learned before we can have a better understanding of the architecture of the *pogo* DNA-protein complex.

8.1.1.4 The activity of *pogo* transposase

The activity of a transposase can be determined by its ability to catalyse transposition reactions including DNA cleavage and strand transfer. For DNA mediated elements which transpose *via* a cut and paste mechanism, the DNA cleavage activity can be tested by an *in vitro* excision assay. In this assay, a supercoiled plasmid DNA containing the transposon is exposed to the transposase in the presence of the divalent ion Mg^{2+} or Mn^{2+}. Since the transposase is supposed to catalyse both 5' and 3' cleavage of the element, the transposon should be excised during the reaction. The DNA cleavage activity of the transposase can therefore be determined by detecting the excised transposon. *Tc1* transposase has been shown to be able to be excise *Tc1* elements out in this way (Vos & Plasterk, 1994). Since *pogo* is structurally similar to *Tc1* and other DNA mediated elements, it is likely to transpose *via* a cut and paste mechanism. The DNA cleavage activity of *pogo* transposase was tested in chapter 7 using the same method as *Tc1*, but no *pogo* transposon was excised out. The supercoiled plasmid was linearised, and the amount of the linearised DNA increased as the reaction time increased. This result indicates that *pogo* transposase failed to catalyse the 5' end, or 3' end, or both ends DNA cleavage reactions in this case. It's probably due to one of the following reasons:

1. The transposase might not be active, and the linearisation of the plasmid is due to the activity of a contaminating DNase activity. The inactivation of the transposase is probably due to one or more of the following reasons:

a). An imperfect DDE motif. Since there is no proper candidate for the last E in the DDE motif of *pogo* transposase, the transposase could have no proper active site to catalyse the 5' cleavage or 3' cleavage or both reactions, so no completely excised transposon could be seen.

b). *pogo* transposase might not be the only protein required for *pogo* transposition. The transposases of *Tc1* and *mariner* have been shown to be able to catalyse transposition reactions in the absence of other proteins (Vos et al., 1996; Dawson, et al., unpublished data). This may explain why they can transpose in such a wide range of hosts. In contrast, *pogo* has been found only restricted to *Drosophila*

melanogaster, indicating it might have very strict requirements for the host, possibly because the transposase is only active in the presence of some host proteins. This is very like the case of *P* elements. *P* elements can only transpose in *D. melanogaster* and a few closely related species, and it has been shown the specific interaction of the *Drosophila* protein IRBP (inverted repeat binding protein) with the specific sequence of the IR of the element is involved in *P* element transposition (Rio & Rubin, 1988). In *E.coli* transposition, *Tn10* requires the host protein IHF while phage *Mu* requires IHF and HU proteins. *pogo* may also need host proteins and/or other factors to cooperate with the transposase to catalyse the transposition reaction.

c). The plasmid pUC18 into which the *pogo* was cloned in chapter 7, might have no proper flanking DNA to enable the two transposon ends and the transposase to assemble to a active synaptic complex, this would prevent the initiation of the transposition reactions.

2. Since there is no direct evidence yet to show that *pogo* is transposed *via* a cut and paste way, the other very unlikely possibility is that *pogo* could be like phage *Mu* and transpose *via* a replicative mechanism. In this case, DNA cleavage would only occur at one end of the transposon, so no completely excised element would be visualised.

3. It could just due to some technical problems which result in the failure in detecting the activity of the transposase.

However, excision assay is only one of the ways to detect the activity of the transposase. Other methods, such as phosphate transfer assay, *in vitro* transposition, and *in vivo* transformation assay (see later) will definitely provide more information about the activity of *pogo* transposase.

8.1.1.5 The regulation of *pogo* transposition

The results presented in this thesis concerning the transposase binding sites and the DNA binding domain of *pogo* transposase together with knowledge of the regulatory mechanism of similar elements such as the *P* element suggest ways in which *pogo* transposition might be regulated:

a) Transposition could be regulated by the specific interaction of transposase with the 5' subterminal transposase binding sites. The transposase binding sites of the 5' subterminal group (position 107-119, 193-202) are located in the region downstream of the TTAA box (position 4) and upstream of the start of the transposase coding region (position 340) (Tudor, et al. 1992). This region might act as the promoter. If the transposase has a lower affinity for the 5' subterminal group transposase binding sites than for the 5' and 3' end group binding sites, transposition

might be regulated by the concentration of the transposase. When the concentration of transposase is low, the transposase only binds to the 5' and 3' end group binding sites, to form the active synaptic DNA-protein complex, so transposition is initiated. When the concentration of transposase get higher, transposase binds to the 5' subterminal binding sites in addition to the 5' and 3' end binding sites. The binding of transposase to the 5' subterminal binding site blocks the transcription, so no more transposase can be synthesised, the concentration of transposase will therefore decrease. In this way, transposase concentration and hence transposition might be negatively regulated by the concentration of the transposase itself.

b) The transposition could be regulated by a repressor protein resulting from alternative splicing of the intron in between the two ORFs. For *P* element, tissue-specific splicing results in the production of the 66 kd repressor protein in somatic cells which inhibits transposition. For *pogo*, if the splicing of the intron in between ORF1 and ORF2 were inhibited, the protein encoded by ORF1 might be made as well as transposase. This might act as a repressor to regulate *pogo* transposition. This ORF1 protein would still have DNA binding ability since it contains the DNA binding domain, but might have no proper catalytic activity and/or dimerisation ability since the C-terminal region of the transposase encoded by ORF2 is missing. This ORF1 protein can therefore compete with the transposase for the DNA binding sites at the ends of the element, but can't catalyse the transposition reactions properly. If transposase monomers have to form multimers to catalyse transposition reactions, then ORF1 protein might form inactive oligomers if it contains the dimerisation domain, or fail to assemble into oligomers if it has no dimerisation domain, therefore block the multimerisation.

8.1.2 CENP-B, transposase, and DNA-protein complex

CENP-B, human centromere protein B, is a 80 kd protein which binds specifically to the 17 bp CENP-B box in centromeric α-satellite DNA (Masumoto, et al., 1989; Muro, et al., 1992). Its amino acid sequence has 51% similarity and 26 % identity with the that of putative *pogo* transposase over almost the entire length of both proteins. The DNA binding domain of CENP-B is located in the N-terminal 125 amino acid region (Yoda, et al., 1992), and the dimerisation domain is located in the C-terminal 59 amino acid region (Kitagawa, et al., 1995). Two CENP-B monomers form a homodimer by their C-terminal dimerisation domains, while each binds to the α-satellite DNA by its N-terminal DNA binding domain. In this way CENP-B might play an important role in the assembly of specific centromere structures by forming the DNA-protein complexes at the sites of CENP-B boxes. This function of CENP-B

makes it very much like a transposase. As an initial step of transposition, transposase monomers also bind specifically to the transposase binding site at each end of the transposon to assemble to a synaptic DNA-protein complex. Moreover, the DNA-protein complex assembled by CENP-B and the synaptic DNA-protein complex assembled by transposase are also very similar, both contains two double stranded DNA and two (or more for transposase) protein monomers. This functional similarity can be explained by their sequence similarity which suggests that CENP-B and transposase could be closely related. CENP-B displays extremely high sequence similarity with *pogo* family transposases. These transposases not only include *pogo* transposase, but also includes those of *Tigger*, the transposon fossil in human genome. Both CENP-B and the *Tigger* transposases are encoded by single exons (Sullivan & Glass, 1991), so the gene coding CENP-B might be even more related to *Tigger*. And CENP-B gene, *Tiggers* and *pogo* might have a common evolutionary origin.

8.2 Future work

In this thesis, the *pogo* putative transposase has been expressed in *E. coli*, the specific DNA binding activity of this transposase has been confirmed, and the DNA binding domain and the transposase binding sites have been isolated. These data make it possible to conduct future studies towards a better understanding of the mechanism of *pogo* transposition.

Four groups of DNA binding sites located at different positions of the element have been isolated. They all have sequences similar to the 12 bp consensus transposase binding site, but the relative affinity of each of the binding site for the transposes is still unknown. These data can been obtained by using same molar amounts of DNA containing different binding sites as probes, and determining the amount of transposase needed when half amount of each probe is bound. If the 5' group and 3' group binding sites have higher affinity than the other two groups, it indicates that the transposase must first bind to the two transposon end to form the synaptic DNA-protein complex; If the transposase binds to sites within each of the end group with different affinities, this might indicate that transposase has to bind to certain sites to form the active DNA-complex; If the binding sites in the 5' subterminal group is bound with lower affinity than those in the two end groups, this will support the hypothesis that these sites could be involved in regulating transposition as described in 7.1.1.5. The relative affinity of transposase for binding sites located in the middle of the transposon will indicate at which stage these binding sites are

As an initial step of transposition, transposase has to bind to the end of the transposon to form the synaptic DNA-protein complex. When the transposon has one transposase binding site at each end, one transposase monomer binds to each end of the transposon, so there must be two transposase monomers in each synaptic DNA-protein complex. When the transposon has more than one transposase binding site at the end, more transposase monomers are probably involved in the complex. The transposase monomers must interact with each other somehow within the complex, and the most likely way is to form a dimer or multimer, as has been shown in the case of *Mu*. Two monomers can be linked with each other by dimerisation. For *pogo*, there are three transposase binding sites at each end of the element, this suggests that there are probably more than two transposase monomers in the DNA-protein complex, the transposase monomers might form an oligomer. In view of this, the ability of transposase monomers to interact with each other should be tested by crosslinking, and the dimerisation domain of the transposase may therefore be identified. It would also be of interset to test how many monomers are involved in each synaptic DNA-protein complex by determining the number of different DNA-transposase complexes formed with increasing amount of the transposase. These data will increase our understanding about how *pogo* transposase interacts with the transposon end sequences, and how the active synaptic DNA-protein complex is assembled.

The DNA binding domain of *pogo* transposase has been shown to be located in the N-terminal 75 amino acid region, a helix-turn-helix motif is also predicted to be present in this region, and its presence is supported by mutagenesis studies described in this thesis. So it is interesting to get a more direct view of the whole structure of the DNA binding motif by structural analysis. The small size of the 75 amino acid DNA binding domain makes it possible to be analysed by nuclear-magnetic-resonance (NMR) analysis. Also since single amino acid substitutions of the positively charged amino acids in the recognition helix abolished the DNA binding ability of the protein; and the introduction of proline in the HTH motif also greatly decreased the DNA binding ability, it would be worthwhile investigating the effect of these changes on the structrure of the DNA binding domain.

It is still very important to study the activity of *pogo* transposase *in vitro* and *in vivo*. In addition to the excision assay described in chapter 7, the *in vitro* activity assay can be conducted in several other ways: a) Phosphoryl transfer assay. This assay is designed to test the ability of transposase to catalyse a strand transfer reaction. In this test, the 3' hydroxyl of a 5'-labelled oligonucleotide attacks a phosphate within another oligonucleotide. If the strand transfer reaction occurs, the labelled oligonucleotide will be transfer to the unlabelled one (Vos & Plasterk, 1994). The phosphoryl transfer

ability of *pogo* transposase could be tested in this way. b) *in vitro* genetic transposition assay. In this assay, a donor plasmid contains the transposon carrying an antibiotic resistance gene, and a target plasmid carrying another antibiotic resistant gene are electroporated into the same *E.coli* cell, and the transposition event can obtained by selecting cells which are double resistant to the antibiotics (Vos, et al., 1996).

The activity of *pogo* transposase could also be tested *in vivo* by microinjection of *pogo* transposase and a plasmid containing a *pogo* transposon into an embryo of a *Drosophila* species containing no *pogo* elements. The transposase activity could be determined if the *pogo* is mobilised in this genome. The ability of *pogo* transposase to function in other species could also be tested in this way.

REFERENCES

Adams, S. E., Mellor, J., Gull, K., Sim, R. B., Tuite, M. F., Kingsman, S. M., and Kingsman, A. J. (1987). The functions and relationship of Ty-VLP proteins in yeasts reflect those of mammalian retroviral proteins. Cell 49, 111-119.

Adzuma, K. and Mizuuchi, K. (1988). Target immunity of *Mu* transposition reflects a differential distribution of Mu B protein. Cell 53, 257-266.

Amemura, J., Ichikawa, H, and Ohtsubo, E. (1990). *Tn3* transposition immunity is conferred by the transposase-binding domain in the terminal inverted-repeat sequence of *Tn3*. Gene 88(1), 21-24.

Allison, R. G., and Chaconas, G. (1992). Role of the A protein-binding sites in the *in vitro* transposition of *Mu* DNA: a complex circuit of interaction involving the *Mu* ends and the transpositional enhancer. J. Biol. Chem. 267, 19963-19970.

Arciszewaka, L. K., Drake, D., and Craig, N. L. (1989). Transposon *Tn7* *cis*-acting sequences in transposition and transposition immunity. J. Mol. Biol. 207, 35-52.

Arciszewaka., K. and Craig, N. L. (1991). Interaction of the *Tn7*-encoded transposition protein TnsB with the ends of the transposon. Nucl. Acids Res. 19, 5021-5029.

Atkinson, P. W., Warren, W. D. and O'Brochta, D. A. (1993). The *hobo* transposable element of *Drosophila* can be cross-mobilized in houseflies and excises like the *Ac* element of maize. Pro. Natl. Acad. Sci. USA 90, 9693-9697.

Augegouillou, C., Bigot, Y., Pollet, N., Hamelin, M. H., Meunierrotival, M., and Periquet, G. (1995). Human and other mammalian genomes contain transposons of the *mariner* family. FEBS Lett. 368 (3), 541-546.

Bainton, R., Gamas, P., and Craig, N. L. (1991). *Tn7* transposition *in vitro* proceeds through and excised transposon intermediate generated by staggered breaks in DNA. Cell 65, 805-816.

Bainton, R. J., Kubo, K. M., Feng, J. N. and Craig, N. L. (1993). *Tn7* transposition: target DNA recognition is mediated by multiple *Tn7*-encoded proteins in a purified *in vitro* system. Cell 72, 931-943.

Baker, T. A., Mizuuchi, M., and Mizuuchi, K. (1991). MuB protein allosterically activates strand transfer by the transposase of phage *Mu*. Cell 65, 1003-1013.

Baker, T. A. and Lou, L. (1994). Identification of residue in the *Mu* transposase essential for catalysis. Proc. Natl. Acad. Sci. USA 91, 6654-6658.

Barth, P. T., Datta, N., Hegges, R. W. and Grinter, N. J. (1976). Transposition of a deoxyribonucleic acid sequence encoding trimethoprim and streptomycin resistances from R483 to other replicons. J. Bacteriol., 125, 800-810.

Bayer, A. A., Lyubomirskaya, N. V., Dzhumagaliev, E. B., Ananiev, E. V., Amiantova, I. G., and Ilyin, Y. V. (1984). Structural orgnaization of transposable element *mdg4* from *Drosophila melanogaster* and a nucleotide sequence of its long terminal repeats. Nucl. Acids. Res. 12(8), 3707-3723.

Beall, E. L., Admon, A., and Rio, D. C. (1994). A *Drosophila* protein homologous to the human p70 Ku autoimmune antigen interacts with the *P* transposable element inverted repeats. Proc. Natl. Acad. Sci. USA 91, 12681-12685.

Belcourt, M. T., and Farabaugh, P. K. (1990). Ribosomal frameshifting in the yeast retrotransposon Ty: tRNA induce slippage on a 7 nucleotide minimal site. Cell 62, 339-352.

Bellen, H. J., O'Kane, C. J., Wilson, C., Grossniklaus, U., Pearson, R. K. and Gehring, W. J. (1989). *P*-element-mediated enhancer detection: a versatile method to study development in *Drosophila*. Genes Dev. 3, 1288-1300.

Benderm J and Kleckner, N. (1986). Genetic evidence that *Tn10* transposes by a nonreplicative mechanism. Cell 45, 801-815.

Benjamin, H. W., and Kleckner, N. (1992). Excision of *Tn10* from the donor site during transposition occurs by flush souble-strand cleavage at the transposon termini. Proc. Natl. Acad. Sci. USA 89, 4648-4652.

Berg, C. A. and Spradling, A. C. (1991). Studies on the rate and site-specificity of *P* element transposition. Genetics 127, 515-524.

Berg, D. E., and How, M. M. (eds), (1989). Mobile DNA. American Society for Microbiology, Washington DC, pp 523-529.

Berg, J. M. (1986). Potential metal-binding domains in nucleic acid binding proteins. Science 232, 485-487.

Berg, J. M. (1990). Zinc finger domains-hypothesis and current knowledge. Annu. Rev. Biophys. Chem. 19, 405-421.

Bier, E., Vaessin, H., Shepherd, S., Lee, K., MdCall, K., Barbel, S., Ackkerman, L., Carretto, R., Uemura, T., Grell, E., Jan, L. Y. and Jan, Y. N. (1989). Searching for pattern and mutation in the *Drosophila* genonme with a *P-lacZ* vector. Genes Dev. 3, 1273-1287.

Bigot, Y., Auge-Gouillou, C. and Periquet, G. (1996). Computer analyses reveal a *hobo*-like element in the nematode *Caenorhabditis elegans*, which presents a conserved transposase domain common with the *Tc1-Mariner* transposon family. Gene 176, 265-271.

Bingham, P. M., Kidwell, M. G., and Rubin, G. M. (1982). The molecular basis of P-M hybrid dysgenesis: the role of the *P* element, a *P* strain-specific transposon family. Cell 29: 995-1004.

Bishia, W. R., Rappuoli, R., and Murphy, J. R. (1987). High-level expression of a proteolytically sensitive diphtheria toxin fragment in *Escherichia coli.* J. Bacteriol. 169, 5140-5151.

Black, D. M., Jackson, M. S., Kidwell, M. G., and Dover, G. A. (1987). KP elements repress *P*-induced hybrid dysgenesis in *Drosophila melanogaster.* EMBO J. 6, 4125-4135.

Blackman, R. K., Grimaila, R., Koehler, M. M. D., and Gelbart, W. M. (1987). Mobilization of *hobo* elements residing within the *decapentaplegic* gene complex: suggestion of a new hybrid dysgenesis system in *Drosophila melanogaster.* Cell 49, 497-505.

Blackman, R. K. and Gelbart, W. M. (1989). The transposable element *hobo* of *Drosophila melanogaster.* In: Berg DE, Howe MM (eds) Mobile DNA. American Society for Microbiology, Washington, D. C., pp 523-529.

Bolland, S. and Kleckner, N. (1995). The two single strand cleavages at each end of *Tn10* occur in a specific order during transposition. Proc. Natl. Acad. Sci. USA 92, 7814-7818.

Bolland, S. and Kleckner, N. (1996). The three chemical steps of *Tn10/IS10* transposition involve repeatedly utilization of a single active site. Cell 84, 223-233.

Boussy, I. A., Charles, L., Hamelim, M. H., Periquet, G. and Shapiro, D. Y. (1993). The occurrence of the transposable element *pogo* in *Drosophila melanogaster.* Genetica 88, 1-10.

Brennan, R. G. (1991). Interactions of the helix-turn-helix binding domain. Curr. Opin. Struct. Biol. 1, 80-88.

Branden, C. and Tooze, J. (1991). Introduction to protein structure. Garland Publishing Co., New York.

Brennan, R. G. (1992). DNA recognition by the helix-turn-helix motif. Curr. Opin. Struct. Biol. 2, 100-108.

Bregliano, J. C. and Kidwell, M. G. (1983). Hybrid dysgenesis determinants. In Mobile Genetic Elements, ed. J. A. Shapiro, pp. 363-410. London: Academic.

Brown, R. S., Sauder, C. and Argos, P. (1985). The primary structure of transcription factor TFIIIA has 12 consencutive repeats. FEBS Lett. 186, 271-274.

Bryan, G. J., Jacobson, J. W., and Hartl, D. L. (1987). Heritable somatic excision of a *Drosophila* transposon. Science 235, 1636-1638.

Bucheton, A., Lavige, J. M., Picard, G., and L'Heritier, P. (1976). Non-mendelian female sterility in *Drosophila melanogaster*: quantitative variations in the efficiency of inducer and reactive strains: Heredity. 36, 305-314.

Bucheton, A., Paro, R., Sang, H. M. Pelisson, A., and Finnegan, D. J. (1984). The molecular basis of *I-R* hybrid dysgenesis in *Drosophila melanogaster*: identification, clonging, and properties of *I* factor. Cell 38, 153-163.

Bucheton, A., Simonelig, M, Vaury, C., and Crozatier, M. (1986). Sequences similar to the *I* transposable element involeved in *I-R* hybrid dysgenesis in *D. melanogaster* occur in other *Drosophila* species. Nature 322, 650-652.

Bujacz, G., Jaskolski, M. Alwxandratos, J., Wlodawer, A., Merkel, G., Katz, R. A., and Skalka, A. M. (1995). High-resolution structure of the catalytic domain of avian sarcoma virus integrase. J. Mol. Biol.235(2), 333.

Calvi, B. R., Hong, T. J., Findley, S. D., and Gelbart, W. M. (1991). Evidence for a common evolutionary origin of inverted repeat transposons in *Drosophila* and plants: *hobo*, *Activator*, and *Tam3*. Cell 66, 465-471.

Calvi, B. R., and Gelbart, W. M. (1994). The basis for germline specificity of the *hobo* transposable element in *Drosophila melanogaster*. EMBO J. 13(7), 1636-1644.

Capy, P., Koga, A., David, J. R., and Hartl, D. L. (1992). Sequence analysis of active *mariner* elements in natrual populations of *Drosophila simulans*. Genetics 130, 499-506.

Chalmers, R. M. and Kleckner, N. (1994). *Tn10/IS10* transposase purification activation, and *in vitro* reaction. J. Biol. Chem. 269, 8029-8035

Chow, S. A., Vincent, K. A., Ellison, V., and Brown, P. O. (1992). Reversal of integration and DNA splicing mediated by integrase of human-immunodeficiency-virus. Science 255, 723-726.

Clark, J. B., Madison, W. P., and Kidwell, M. G. (1994). Phylogenetic analysis supports horizontal transfer of *P* transposable elements. Mol. Biol. Evol. 11(1), 40-50.

Coen, E. S., Carpenter, R., and Martin, C. (1986). Transposable elements generate novel spatial pattern of gene expression in *A. majus*. Cell 47, 285-296.

Collins, J. J. and Anderson, P. (1994). The *Tc5* family of transposable elements in *Caenorhabditis elegans*. Genetics 137, 771-781.

Colloms, S. D., van Luenen, H. G. A. M., and Plasterk, R. H. A. (1994). DNA binding activities of the *Caenorhabditis elegans Tc3* transposase. Nucl. Acids Res. 22(5), 5548-5554.

Cooley, L., Berg, C., and Spradling, A. (1988a). Controlling *P* element insertional mutagenesis. Trends Genet. 4, 254-258.

Cooly, L., Delley, R. and Spradling, A. (1988b). Insertional mutagenesis of the *Drosophila* genome with single *P* elements. Science 239, 1121-1128.

Craig, N. L. (1995). Unity in transpositon reactions. Science 270, 253-254.

Craig, N. L. (1996). Transposon *Tn7*. Curr. Top. Microbiol. Immunol. 204, 27-48.

Craigie, R., Mizuuchi, M., and Mizuuchi, K. (1984). Site-specific recognition of the bacteriophage *Mu* ends by the Mu A protein. Cell 39, 387-394.

Craigie, R., Fujiwara, T., and Bushman, F. (1990). The IN protein of moloney murine leukemia virus processes the viral DNA ends and accomplishes their integration *in vitro*. Cell 62, 829-837.

Daboussi, M. J., Langin, T., and Brygoo, Y. (1992). *Fot1*, a new family of fungal transposable elements. Mol. Gen. Genet. 232, 12-16.

Daniels, S. B., Petersoon, K. R., Strausbaugh, L. D., Kidwell, M, G. and Chovnick, A. (1990). Evidence for horizontal transmission of *P* transposable element between *Drosophila* species. Genetics 124, 339-355.

Daniels, S. B., Chovnick, A., and Boussy, I. A. (1990). Distribution of *hobo* transposable elements in the genus *Drosophila*. Mol. Biol. Evol. 7, 589-606.

Dawson, A., Hartswood, E., Paterson, T. and Finnegan, D. J. (1997). A LINE-like element in *Drosophila*, the *I* factor, encodes a protein with properties similar to those of retroviral nucleocapsids. EMBO J. (to be published).

Derbyshire, V., Grindley, N. D. F., and Joyce, C. M. (1991). The 3'-5' exonuclease of DNA polymerase I of *Escherichia coli*: contribution of each amino acid at the active site to the reaction. EMBO J. 10, 17-24.

Di Nocera, P. P., and Casari, G (1987). Related polypeptides are encoded by *Drosophila F* element, *I* factors and mammalian *L1* sequences. Proc. Natl. Acak. Sci. USA 84, 5843-5847.

Doak, T. G., Doerder, T. P., Jahn, C. L., and Herrick, G. (1994). A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. Proc. Natl. Acad. Sci. USA 91, 942-946.

Driver, A., Lacey, S. F., Cullingford, T. E., and O'Hare, K. (1989). Structural analysis of *Doc* transposable elements associated with mutation at the *white* and *suppressor of forked loci* of *Drosophila melanogaster*. Mol. Gen. Genet. 220, 49-52.

Dyda, F., Hickman, A. B., Jenkins, T. M., Engelman, A., Craigie, T., and Davies, D. R. (1994). Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. Science 266, 1981-1986.

Earnshaw, W. C., Sullivan, K. T., Machlin, P. S., Cooke, C. A., Kaiser, D. A., Pollard, T. D., Rothfield, N. F., and Cleveland, D. W. (1987). Molecular clonging of cDNA for CENP-B, the major human centromere autoantigen. J. Cell Biol. 104, 817-829.

Engelman, A. and Craigie, R. (1992). Identification of conserved amino acid residues critial for human immunodeficiency virus type 1 integrase function *in vitro*. J. Virol. 66, 6361-6369.

Engels, W. R. (1983). The *P* family of transposable elements in *Drosophila*. Annu. Rev. Genet. 17, 315-344.

Engels, W. R. (1989). *P* elements in *Drosophila melanogaster*. In: Berg DE, Howe MM (eds) Mobile DNA. American Society for Microbiology, Washington, D. C., pp 437-484.

Eichinger, D. J., and Boeke, J. D. (1988). The DNA intermediate in yeast *Ty1* element transposition copurifies with virus-like particles : cell free *Ty1* transposition. Cell 54, 955-966.

Fawcett, D. H., Lister, C. K., Kellett, E., Finnegan, D. J. (1986). Transposable elements controlling I-R hybrid dysgenesis in *D. melanogaster* are similar to mammaliuan LINEs. Cell 47, 1007-1015.

Fayet, O., Ramond, P., Polard, P., Prere, M. F., and Chandler, M. (1990). Functional similarities between retroviruses and the *IS3* family of bacterial insertion sequences. Mol. Microbiol. 4, 1771-1777.

Feldmar, S. and Kunze, R. (1991). The ORFa protein, the putative transposase of maize transposable element *Ac*, has a basic DNA binding domain. EMBO J. 10, 4003-4010.

Finnegan, D. J. (1992). Transposable elements. Curr. Opin. Genet. Dev. 2, 861-867.

Freund, R., and Meselson, M. (1984). Long terminal repeat nucleotide sequence and specific insertion of the *gypsy* transposon. Proc. Natl. Acad. Sci. USA 81, 4462-4464.

Fried, M. and Crothers, D. M. (1981). Equilibria and kinetics of *lac* repressor-operator interactions by polyacrylamide gel electrophoresis. Nucl. acids. Res. 9, 6505-6525.

Fried, M. and Crothers, D. M. (1983). CAP and RNA polymerase interactions with the *lac* promoter: binding stoichiometry and long range effects. Nucl. Acids Res. 11, 141-148.

Fried, M. and Crothers, D. M. (1984). Kinetics and mechanism in the reaction of gene regulatory proteins with DNA. J. Mol. Biol. 172, 263-282.

Fusswinkel, H., Schein, S., Coruage, U., Starlinger, P., and Kunze, R. (1991). Detection and abundance of mRNA and protein encoded by transposable element *Activator* (*Ac*) in maize Mol. Gen. Genet. 225, 186-192.

Garcia-Femandez, J., Marfany, G., Baguna, J., and Salo, T. (1993). Infiltration of *mariner* elements. Nature 364, 109.

Garfinkel, D. J., Boeke, J. D., and Fink, G. R. (1985). Ty element transposition: reverse transcriptase and virus-like particles. Cell 42, 507-517.

Garfinkel, D. F., Hedge, A. M., Youngren, S. D., and Sanders, N. J. (1991). Proteolytic processing of *pol*-TYB proteins from the yeast retrotransposon *Ty1*. J. Virol. 65, 4573-4581.

Gearing, D. P., Nicola, N. A., Metcalf, D., Foote, S., Willson, T. A., Gough, N. M., and Williams, R. L. (1989). Production of leukemia factor in *Escherichia coli* by a novel procedure and its use in maintaining embryonic stem cells in culture. BioTechnology 7, 1157-1161.

Germino, J. and Bastia, D. (1984). Rapid purification of a gene product by genetic fusion and site-specific proteolysis. Proc. Natl. Acad. Sci. USA 81, 4692-4696.

Gierl, A., Lutticke, S., and Saedler, H. (1988). TnpA product encoded by the transposable element *En1* of *Zea mays* is a DNA binding protein. EMBO J. 7, 4045-4053

Grindley, N. D. G. and Leschziner, A. E. (1995). DNA transposition: from a black box to a color monitor. Cell, 83, 1063-1066.

Haffey, M. L., Lehman, D., and Boger, J. (1987). Site-specific cleavage of a fusion protein by renin. DNA 6, 565-571.

Hagemann, S., Miller, W. J., and Pinsker, W. (1992). Identification of a complete *P*-element in the genome of *Drosophila bifasciata*. Nucl. Acids Res. 20, 409-413.

Haniford, D. B., Chelouche, A. and Kleckner, N. (1989). A specific class of *IS10* transposase mutants are blocked for target site interactions and promote formation of an excised transposon fragment. Cell 59, 385-394.

Haniford, D. B., Benjamin, H. W. and Kleckner, N. (1991). Kinetic and structrual analysis of a cleaved donor intermediae and a strand transfer intermediate in *Tn10* transposition. Cell 64, 171-179.

Haniford, D. B. and Kleckner, N. (1994). *Tn10* transposition *in vivo*: temporal separation of cleavages at the two transposon ends and roles of terminal base pairs subsequent to interaction of ends. EMBO J. 13, 5847-5851.

Harris, L. J., Baillie, D. L., and Rose, A. M. (1988). Sequence identity between an inverted repeat family of transposable elements in *Drosophila* and *Caenorhabditis*. Nucl. Acids Res. 16, 5591-5598.

Harrison, S. C. and Aggarwal, A. K. (1990). DNA recognition by proteins with the helix-turn-helix motif. Annu. Rev. Biochem. 59, 933-969.

Harrison, S. C. (1991). A structural taxonomy of DNA-binding domains. Nature 353, 715-719.

Hartl, D. (1989). Transposable element *mariner* in *Drosophila* species. In: Berg DE, Howe MM (eds) Mobile DNA. American Society for Microbiology, Washington, D. C., pp 531-536.

Haymer, D. S. and Marsh, J. L. (1986). Germline and somatic instability of a *white* mutation in *Drosophila mauritiana* due to a transposable element. Dev. Genet. 6, 281-291.

Hehl, R., Nacken, W. K. F., Krause, A., Saedler, H., and Sommer, H. (1991). Structural analysis of *Tam3*, a transposable element from *Antirrhinum majus*, reveals homologies to the *Ac* element from maize. Plant Mol. Biol. 16, 369-371.

Henikoff, S. (1992). Detection of *Caenorhabditis* transposon homologs in divers organisms. New Biol. 4, 382-388.

Henikoff, A. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl. Acad. Sci. USA 89, 10915-10919.

Hohmann, S. (1993). Characterisation of *PDC2*, a gene necessary for high level expression of pyruvate decarboxylase structural genes in *Saccaromyces cerevisiae*. Mol. Gen. Genet. 241, 657-666.

Hope, I. A., and Struhl, K. (1986). Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of yeast. Cell 46, 885-894.

Huisman, O., Errada, P. R., Signon, L., and Kleckner, N. (1989). Mutational analysis of *IS10* outside end. EMBO J., 8, 2101-2109.

Ichikawa, H., Ikeda, K., Wishart, W. L., and Ohtsubo, E. (1987). Specific binding of transposase to terminal inverted repeats of transposable element *Tn3*. Proc. Natl. Acad. Sci. USA 84, 8220-8224.

Ichikawa, H., Ikeda, D., Amemura, J., and Ohtsubo, E. (1990). Two domains in the terminal inverted-repeat sequence of transposon *Tn3*. Gene 86, 11-17.

Itakura, K., Hirose, T., Crea, R. M. Riggs, A. D., Heyneker, H. L., Bolivar, F., and Boyer, H. W. (1977). Expression in *E. coli* of a chemically synthesized gene for the horomone somatostatin. Science 198, 1053-1056.

Jacobson, J. W. and Hartl, D. L. (1985). Coupled instability of two X linked genes in *Drosophila mauritiana*: Germinal and somatic mutability. Genetics 111, 57-65.

Jacobson, J. W., Medhora, M. M., and Hartl, D. L. (1986). Molecular structure of a somatically unstable transposable element in *Drosophila*. Proc. Natl. Acad. Sci. USA 83, 8684-8688.

Jakubczak, J. L., Xiong, Y., and Eickbush, T. H. (1990). Type I (R1) and type II (R2) ribosomal DNA insertions of *Drosophila melanogaster* are retrotransposable elements closely related to those of *Bombix mori*. J. Mol. Biol. 212, 37-52.

Johnson, M. S., McClure, M. A., Feng, D. F., Gray, J., and Doolittle, R. F. (1986). Computer analysis of retroviral *pol* genes: assignment of enzymatic functions to specific sequences and homologies with nonviral enzymes. Proc. Natl. Acad. Sci. USA 83, 7648-7652.

Kachroo, P., Leong, S. A. and Chattoo, B. B. (1994). *Pot2*, an inverted repeat transposon from the rice blast fungus *Magnaporthe grisea*. Mol. Gen. Genet. 245, 339-348.

Kaptein, R. (1991). Curr. Opin, Struc. Biol. 1, 63-70.

Karess, R. E. and Rubin, G. M. (1984). Analysis of *P* transposable element functions in *Drosophila*. Cell 38, 135-146.

Katz, R. A., Merkel, J., Kulkosky, J., Leis, J., and Skalka, A. M. (1990). The avian retroviral IN protein is both necessary and sufficient for integrative recombination *in vitro*. Cell 63, 87-95.

Katz, R. A., and Skalka, A. M. (1994). The retroviral enzymes. Annu. Rev. Biochem. 63, 133-173.

Kaufman, O. K., Doll, R. F. and Rio, D. C. (1989). *Drosophila P* element transposase recognizes internal *P* element DNA sequences. Cell 59, 359-371.

Kidwell, M. G. (1986). *P-M* mutagenesis, pp. 59-81 in *Drosophila*: a practical approach, Edited by D. B. Roterts. IRL Press, Oxford.

Kim, A., Terzian, C., Santamaria, P., Pelisson, A., Prud'homme, N., and Bucheton, A. (1994). Retroviruses in invertebrates: The *gypsy* retrotransposon

is apparantly an infectious retrovirus of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 91, 1285-1289.

Kim, K., Namgoong, S. Y., Jayaram, M. and Harshey, R. M. (1995). Step-arrest mutants of phage *Mu* transposase. J. Biol. Chem. 270(3), 1472-1479.

Kimmel, B. E., Moioi, O. K., and Young, J. R. (1987). *Ingi*, a 5.2-kilobase dispersed sequence element from *Trypanosoma brucei* that carries half of smaller mobile element at either end and has homology with mammalian LINEs. Mol. Cell. Biol. 7, 1465-1475.

Kitagawa, K., Masumoto, H., Ikeda, M., and Okazaki, T. (1995). Analysis of protein-DNA and protein-protein interactions of centromere protein B(CENP-B) and properties of the DNA-CENP-B complex in the cell cycle. Mol. Cell. Biol. 15(3), 1602-1612.

Kleckner. N. (1989). Transposon *Tn10*. In Berg, D. E. and Howe, M. M. (eds), Mobile DNA. American Society for Micobiology, Washington, DC, pp. 335-374.

Kleckner, N., Chalmers, R., Kwon, D., Sakai, J. and Bolland, S. (1995). *Tn10* and *IS10* transposition and chromosome rearrangements: mechanism and regulation *in vivo* and *in vitro*. Curr. Top. Microbiol. Immunol. 204, 49-82.

Klug, A., and Rhodes D. (1987). 'Zinc fingers': a novel protein motif for nucleic acid recognition. Trends Biochem. Sci. 12, 464-469.

Kubo, K. M. and Craig, N. L. (1990). Bacterial transposon *Tn7* utilizes two classes of target sites. J. Bacteriol. 172, 2774-2778.

Kulkosky, J., Jones, K. S., Katz, R. A., Mack, J. P. G. and Skalka, A. M. (1992). Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases. Mol. Cell. Biol. 12(5): 2331-2338.

Kunze, R. and Starlinger, P. (1989). The putative transposase of transposable element *Ac* from *Zea mays* L. interacts with subterminal sequences of *Ac*. EMBO J. 8(11), 3177-3185.

Kunze, R., Behrens, U., Coruage, F. U., Feldmar, S., Kuhn, S. and Lutticke, R. (1993). Dominant transposition-deficient mutants of maize *Activator* (*Ac*) transposase. Proce. Natl. Acad. Sci. USA 90, 7094-7098.

Lachaise, D., Kariou, M., David, J. R., Lemurnier, F., Tsacas, L., and Ashburner, M. (1988). Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. 22, 159-227.

Lampe, D. J., Churchill, M. E. A., and Robertson, H. M. (1996). A purified *mariner* trasposase is sufficient to mediate transposition *in vitro*. EMBO J. 15(9), 5470-5479.

Landschulz, W. H., Johnson, P. F., and McKnight, S. L. (1988). The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. Science 240, 1759-1764.

Lansman, R. A., O'Shad, R. O., Girigiatti, T. A., and Brodk, H. W. (1987). Evolution of *P* transposable elements: sequences of *Drosophila nebulosa P* elements. Proc. Natl. Acad. Sci. USA 84, 6491-6495.

Laski, F. A., Rio, D. C., and Rubin, G. M. (1986). Tissue specificity of *Drosophila P* element transposition is regulated at the level of mRNA splicing. Cell 44, 7-19.

Lavoie, B. C. and Chaconas, G. (1990). Immunoelectron microscopic analysis of the A, B, and HU protein content of bacteriophage *Mu* transpososomes. J. Biol. Chem. 265, 1623-1627.

Lee, C. C., Mul, Y. M. and Rio, D. C. (1996). The *Drosophila P*-element KP repressor protein dimerizes and interacts with multiple sites on *P*-element DNA. Mol. Cell. Biol. 16(10): 5616-5622.

Lee, N., Cozzikorto, J., Wainwright, N., and Testa, D. (1984). Cloning with tandem gene systems for high level gene expression. Nucl. Acids Res. 12, 6797-6812.

Leung, P. C., Teplow, D. B., and Harshey, R. B. (1989). Interaction of distinct domains in *Mu* transposase with *Mu* DNA ends and an internal transpositional enhancer. Nature 338, 656-658.

Li, M., and Starlinger, P. (1990). Mutational analysis of the N terminus of the protein of maize transposable element *Ac*. Proc. Natl. Acad. Sci. USA 87, 6044-6048.

Li, W. and Shaw, J. E. (1993). A variant *Tc4* transposable element in the nematode *C. elegans* could encode a novel protein. Nucl. Acids Res. 21(1), 59-67.

Lim, J. K. (1989). Intrachromosomal rearrangements mediated by *hobo* transposons in *Drosophila melanogaster*. Proc.Natl. Acad. Sci. USA 85, 9153-9157.

Marlor, R., Parkhurst, S. M. and Corces, V. R. (1986). The *Drosophila melanogaster Gypsy* transposable element encodes putative gene products homologous to retroviral proteins. Mol. Cell Biol. 6(4), 1129-1134.

Maruuama, K. and Hartl, D. L. (1991). Evolution of transposable element *mariner* in *Drosophila* species. Genetics 128, 319-329.

Maruyama, K., Schoor, K. D., and Hartl, D. L. (1991). Identification of nucleotide substitutions necessary for *trans*-activation of *mariner* transposable elements in *Drosophila*: analysis of maturally occuring elements. Genetics 128, 777-784.

Masumoto, H., Masukata, H., Muro, Y., Nozaki, N., and Okazaki, T. (1989). A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. J. Cell Biol. 109, 1963-1973.

Mathias, S. L., Scott, A. F., Kazazian, H. H., Boeke, J. D. and Gabriel, A. (1991). Reverse-transcriptase encoded by a human transposable element. Science 254, 1808-1810.

May, E. W. and Craig, N. L. (1996). Switching from cut-and-paste to replicative *Tn7* transposition. Science 272, 401-404.

Maxwell, A., Craigie, R., and Mizuuchi, K. (1987). B protien of bacteriophage *Mu* is an ATPase that preferentially stimulates intermolecular DNA strand transfer. Proc. Natl. Acad. Sci. USA 84, 699-703.

McCintock B. (1949). Mutable loci in maize. Carnegie Inst Washington Year book 48, 142-154.

McCintock B. (1951). Chromosome organization and gene expression. Cold Spring Harbor Symp Quant Biol. 16, 13-47.

McCintock B. (1965). The control of gene action in maize. Brookhaven Symp. Biol. 18, 162-184.

Mcginnis, W., Shermoen, A. W., and Beckendorg, S. K. (1983). A transposable element inserted just 5' to a *Drosophila* glue protein gene alters gene expression and chromatin structrue. Cell 34, 75-84.

Medhora, M. M., Mauyama, K., and Hartl, D. L. (1991). Molecular and functional analysis of the *mariner* muator element *Mos1* in *Drosophila*. Genetics 128, 311-318.

Mellor, J., Malim, M., Gull, K., Tuite, M. F., McCreedy, S., Dibbayawan, T., Kingsman, S. M., and Kingsman, J. (1985). Reverse transcriptase activity and Ty RNA are associated with virus-like particles in yeast. Nature 318, 583-586.

Miller, J., McLachian, A. D., and Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. EMBO J. 4, 1609-1614.

Miller, W. J., Hagemann, S., Reiter, E., and Pinsder, W. (1992). *P*-element homologous sequences are tandemly repeated in the genome of *Drosophila guanche*. Proc. Natl. Acad. Sci. USA 89, 4018-4022.

Misra, S. and Rio, D. C. (1990). Cytotype control of *Drosophila P* element transposition: the 66 kd protein is a repressor of transposase activity. Cell 62, 269-284.

Mizuuchi, M., and Mizuuchi, K. (1989). Efficient *Mu* transposition requires interaction of transposase with a DNA sequence at the *Mu* operator: implications for regulation. Cell 58, 399-408.

Mizuuchi, M., Baker, T. A., and Mizuuchi, K. (1991). DNase protection analysis of the stable synaptic complexes involved in *Mu* transposition. Proc. Natl. Acad. Sci. USA 88, 9031-9035.

Mizuuchi, M., Baker, T. A., and Mizuuchi, K. (1992). Assembly of the active form of the transposase-*Mu* DNA complex: a critical control point in *Mu* transposition. Cell 70, 303-311.

Mizuuchi, K. (1992). Transpositional recombination: mechanistic insight from studies of *Mu* and other elements. Annu. Rev. Biochem. 61, 1011-1051.

Moerman, D. G. and Waterson, R. H. (1989). Mobile elements in *Caenorhabditis elegans* and other nematodes. In: Berg DE, Howe MM (eds) Mobile DNA. American Society for Microbiology, Washington, D. C., pp 537-556.

Morgan, G. T. (1995). Identification in the human genome of mobile elements spread by DNA-mediated transposition. J. Mol. Biol. 254 (1), 1-5.

Mori, I., Moermann, D. G., and Waterston, R. H. (1988). Analysis of a mutator activity necessary for germline transposition and excision of *Tc1* transposable elements in *Caenorhabditis elegans*. Genteics 120, 397-407.

Mullins, M. C., Rio, D. C., and Rubin, G. M. (1989). Cis-acting DNA sequence requirements for *P* element transpositoon. Genes Dev. 3, 729-738.

Muller, F., Bruhler, K. H., Freidel, K., Kowallik, K. V., and Ciriacy, M. (1987). Processing of *Ty1* proteins and formation of *Ty1* virus-like particles in *Saccharomyces cerevisiae*. Mol. Gen. Genet. 207, 421-429.

Muller-Neumann, M., Yoder, J. I., and Starlinger, P. (1984). The DNA sequence of the transposable element *Ac* of *Zea mays* L. Mol. Gen. Genet. 198, 19-24.

Muro, Y., Masumoto, H., Yoda, K., Nozaki, N., Ohashi, M., and Okazaki, T. (1992). Centromere protein B assembles human centromeric α-satellite DNA at the 17 bp sequence, CENP-B box. J. Cell Biol. 116, 585-596.

Murre, C., McCaw, P. S., and Baltimore, D. (1989). A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, myoD, and myC proteins. Cell 56, 777-783.

Murre, C., McCaw, P. S., Vaessin, H., Caudy, M., and Jan, L. Y. (1989). Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. Cell 58, 537-544.

Nagai, K. and Thogersen, H. C. (1984). Generation of β-globin by sequence-specific proteolysis of a hybrid protein produced in *Escherichia coli.* Nature (London) 309, 810-812.

O'Hare, K. and Rubin, G. M. (1983). Structure of *P* transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. Cell 34, 25-35.

O'Hare, K., Alley, M. R., Cullingford, T. E., Driver, A., and Sanderson, M. J. (1991). DNA sequence of the *Doc* retroposon in the *white-one* mutant of *Drosophila melanogaster* and of secondary insertions in the phenotipically altered derivatives *whith-honey* and *white-eosin.* Mol. Gen. Genet. 225, 17-24.

O'Kane, C. J., and Gehring, W. J. (1987). Detection *in situ* of genomic regulatory elements in *Drosophila.* Proc. Natl. Acad. Sci. USA 84, 9123-9127.

Oosumi, T., Belknap, W. R., and Garlick, B. (1995). *Mariner* transposons in humans. Nature 378 (6558), 672.

Pabo, C. O. and Sauer, R. T. (1984). Protein-DNA recognition. Annu. Rev. Biochem. 53, 293-321.

Pabo, C. O. and Sauer, R. T. (1992). Transcription factors: structural families and principles of DNA recognition. Annu. Rev. Biochem. 61, 1053-1059.

Panaccione, D. G., Pitkin, J. W., and Walton, J. D., and Annis, S. L. (1996). Transposon-like sequencex at the *TOX2* locus of the plant-pathogenic fungus *Cochliobolus carbonum.* Gene 176, 103-109.

Paricio, N., Petez-Alongo, M., Martinez-Sebastian, M. J., and Frutos, R. (1992). *P* sequences of *Drosophila subobscura* lack exon 3 and may encode a 66 kd repressor-like protein. Nucl. Acids Res. 19, 6713-6718.

Pietrokovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. Nucl. Acids Res. 24(19): 3836-3845.

Pietrokovski, S. and Henikoff, S. (1997). A helix-turn-helix DNA-binding motif predicted for *Tc1*, *mariner* and *pogo* family transposases. Mol. Gen. Genet. (to be publised).

Plasterk, R. H. A. (1987). Differences between *Tc1* elments from the *C. elegans* strains. Nucl. Acids. Res. 15, 10050.

Pohlman, R. F., Fedoroff, N. V., and Messing, J. (1984). The nucleotide sequence of the maize controlling element *Activator.* Cell 37, 635-643.

Polard, P. and Chandler, M. (1995). Bacterial transposases and retroviral integrases. Mol. Micro. 15(1), 13-23.

Prendergast, G. C., and Ziff, E. B. (1989). DNA-binding motif. Nature 341, 392.

Priimagi, A. F., Mizrokhi, L. J., and Ilyin, Y. Y. (1988). The *Drosophila* mobile element *jockey* belongs to LINEs and contains coding sequences homologous to some retroviral proteins. Gene 70, 253-262.

Radstrom, P., Skold, O., Swedberg, G., Flensburg, F., Roy, P. H., and Sundstrom, S. (1994). Transposon *Tn 5090* of plasmid R751, which carries an integron, is related to *Tn7*, *Mu*, and the retroelements. J. Bacteriol. 176, 3257-3268.

Rasmusson, D. E., Raymond, J. D., and Simmons, M. J. (1993). Repression of hybrid dysgenesis in *Drosophila melanogaster* by individual naturally occurring *P* elements. Genetics 133, 605-622.

Revzin, A., Cehlarek, J., and Garner, M. (1986). Comparison of nucleic-acid interactions in solution and in polyacrylamide gel. Anal. Biochem. 153, 172-177.

Rice, R., and Mizuuchi, K. (1995). Structure of bacteriophage *Mu* transposase core: a common structural motif for DNA transposition and retroviral integration. Cell 82, 209-220.

Rio, D. C., Laski, F. A., and Rubin, G. M. (1986). Identification and purification of a *Drosophila* protein that binds to the terminal 31-base-pair inverted repeats of the *P* transposable element. Proc. Natl. Acad. Sci. USA 85, 9829-8933.

Rio, D. C., and Rubin, G. M. (1988). Identification and purification of a *Drosophila* protein that binds to the terminal 31-base-pair inverted repeats of the *P* transposable element. Proc. Natl. Acad. Sci. USA 85, 8929-8933.

Rio, D. C. (1990). Molecular mechanisms regulating *Drosophila P* element transposition. Annu. Rev. Genet. 24, 543-578.

Robertson, H. M. and Engels, W. R. (1989). Modified *P* elements that mimic the *P* cytotype in *Drosophila melanogaster*. Genetics 123, 815-824.

Robertson, H. M. (1993). The *mariner* element is widespread in insects. Nature 362, 241-245.

Robertson, H. M. (1995). The *Tc1-mariner* superfamily of transposons in animals. J. Insect Physiol. 41(2), 99-105.

Robertson, H. M. (1996). Members of the *pogo* superfamily of DNA-mediated transposons in the human genome. Mol. Gen. Genet. 252, 761-766.

Robertson, H. M. and Asplund, M. L. (1996). *Bmmar1*: a basal lineage of the *mariner* family of transposable elements in the silkworm moth, *Bombyx mori*. Insect biochem. Mol. Biol. 26 (8-9), 945-954.

Rogers, M., Ekaterinaki, N., Nimmo, E. and Sherratt, D. (1986). Analysis of *Tn7* transposition. Mol. Gen. Genet. 205, 550-556.

Rosenzweig, B., Liao, L. W., and Hirsch, D. (1983). Sequence of the *C. elegans* transposable element *Tc1*. Nucl. Acids Res.12, 4201-4209.

Rowland, S, J. and Dyke, K. G. H. (1990). *Tn522*, a novel transposable element from *Staphylococcus aureus*. Mol. Microbiol. 4, 961-965.

Rubin, G. M. and Spradling, A. C. (1982). Genetic transformation of *Drosophila* with transposable element vectors. Science 218, 348-353.

Rubin, G. M., Kidwell, M. G., Bingham, P. M. (1982). The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. Cell 29, 987-994.

Ruvolo, V., Hill, J. E., and Levitt, A. (1992). The *Tc2* transposon of *Caenorhabditis-elegnas* has the structure of a self-regulated element. DNA Cell Biol. 11(2): 111-122.

Sakai, J., Chalmers, R. H. and Kleckner, N. (1995). Identification and characterization of a pre-cleavage synaptic complex that is an early intermediate in *Tn10* transposition. EMBO J. 14, 4374-4383.

Sarnovsky, R. J., May, E. W., and Craig, N. L. (1996). The *Tn7* transposase is a heteromeric complex in which DNA breakage and joining activities are distributed between different gene products. EMBO J. 15(22), 6348-6361.

Schein, C. H. (1989). Production of soluble recombinant proteins in bacteria. BioTechnology 7, 1141-1149.

Sezutsu, H., Nitasaka, E., and Yamazaki, T. (1995). Evolution of the LINE-like *I*-element in the *Drosophila melanogaster*. Mol. Gen. Genet. 249, 168-178.

Signon, L., and Kleckner, N. (1995). Negative and positiove regulation of *Tn10/IS10*-promoted recombination by IHF: two distinguishable processes inhibit transposition off multicopy plasmid replicons and activate chromosomal events that favor evolution of new transposons. Genes Dev. 9, 1123-1136.

Simmons, M. J., Raymond, J. D., Rasmusson, K. E., Miller, L. M., McLarnon, C. F., and Zunt, J. R. (1990). Repression of *P* element-mediated hybrid dysgenesis in *Drosophila melanogaster*. Genetics 124, 663-676.

Simmons, G. M. (1992). Horizontal transfer of *hobo* transposable elements within the *Drosophila melanogaster* species complex: evidence from DNA sequencing. Mol. Biol. Evol. 9(6), 1050-1060.

Simonelig, M., and Anxolabehere, A. (1991). A *P* element of *Scaptomyza pallida* is active in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 88, 6102-6106.

Smit A. F. A. and Riggs, A. D. (1996). *Tiggers* and other DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci. USA 93, 1443-1448.

Smith, D. B. and John, K. S. (1988). Single-step purification of polypeptides expresed in *Escherichia coli* as fusions with glutathion S-transferase. Gene 67, 31-40.

Smith, D., Wohlgemuth, K., Calvi, B. R., Franklin, I., and Gelbart, W. M. (1993). *hobo* enhancer trapping mutagenesis in *Drosophila* reveal an insertion specificity differert from *P* elements. Genetics 135, 1063-1076.

Smith, J. (1997). Ph. D thesis.

Sommer, H., Carpenter, R., Harison, B. J.. Saedler, H. (1985). The transposable element *Tam3* of *A. majus* generates a novel type of sequence alterations upon excision. Mol. Gen. Genet. 199, 225-231.

Song, S. U., Gerasiova, T., Kurkulos, M, Boeke, J. D., and Corces, V. G. (1994). An *Env*-like protein encoded by a *Drosophila* retroelement: evidence that *gypsy* is an infectious retrovirus. Genes Dev. 8, 2046-2057.

Spradling, A. C. (1986). *P* element-mediated transformation. In *Drosophila*: a Prectical Approch, ed. D. B. Roberts, pp. 175-197. Oxford: IRL press.

Stalder, R., Caspers, P. Olasz, F., and Arbert, W. (1990). The N-terminal domain of the insertion sequence 30 transposase interacts specifically with the terminal inverted repeats of the element. J. Biol. Chem. 265 (7), 3757-3762.

Streck, R. D., MacGaffey, J. E., and Beckendrof, S. K. (1986). The structure of *hobo* transposable elements and their insertion site. EMBO J. 5(13), 3615-3623.

Sullivan, K. F. and Glass, C. A. (1991). CENP-B is a highly conserved mammalian centromere protein with homology to the helix-loop-helix family of proteins. Chromosoma 100, 360-370.

Surette, M. G., Buch, S. J., and Chaconas, G. (1987). Transpososome: stable protein-DNA complexes involved in the *in vitro* transpositon of bacteriophage *Mu* DNA. Cell 49, 253-262.

Surette, M. G., Lavoie, B. D., and Chaconas, G. (1989). Action at a distance in *Mu* DNA transposition: an enhancer-like element is the site of action of supercoiling relief activity by integartion host factor (IHF). EMBO J. 8, 3483-3489.

Szoka, P. R., Schreiber, A. B., Chan, H., and Murthy, J. (1986). A general method for retrieving the components of a gentically enginered fusion protein. DNA 5, 11-20.

Toth, M., Grimsby, J., Buzsaki, G., and Donovan, G. P. (1995). Epileptic seizures caused by inactivation of a novel gene, *jerky*, related to centromere binding protein-B in transgenic mice. Nature Genetics 11, 71-75.

Trentmann, S. M., Saedler, H., and Gierl, A. (1993). The transposable element *En/Spm*-encoded TNPA protein contains a DNA binding and a dimerization domain. Mol. Gen. Genet. 238: 201-208.

Tudor, M., Lobocka, M., Goodell, M., Pettitt, J., and O'Hare, K. (1992). The *pogo* transposable element family of *Drosophila melanogaster*. Mol. Gen. Genet. 232, 126-134.

van Gent, D. C., Oude Groeneger, A. A. M. and Plasterk, R. H. A. (1992). Mutational analysis of the integrase protein of human immunodeficiency virus type 2. Proc. Natl. Acad. Sci. USA 89, 9598-9602.

van Luenen, H. G. A. M., Colloms, S. D., and Plasterk, R. H. A. (1993). Mobilization of quiet, endogenous *Tc3* transposons of *Caenorhabditis elegans* by forced expression of *Tc3* transposase. EMBO J. 12, 2513-2520.

van Luenen, H. G. A. M. and Plasterk, R. H. A. (1994). Target site choice of the related transposable elements *Tc1* and *Tc3* of *Caenorhabditis elegans*. Nucl. Acids Res. 22(3), 262-269.

van Luenen, H. G. A. M., Colloms, S. D., and Plasterk, R. H. A. (1994). The mechanism of transposition of *Tc3* in *C. elegans*. Cell 79, 293-301.

Villa, S., De Fazio, G., Donini. S., Tarchi. G., and Canosi, U. (1988). Expression in *Escherichia coli* and characterization of human growth-hormone-releasing factor. Eur. J. Biochem. 171, 137-141.

Voronova, A., and Baltimore, D. (1990). Mutations that disrupt DNA binding and dimer formation in the E47 helix-loop-helix protein map to distinct domains. Proc. Natl. Acad. Sci. USA 87, 4722-4726.

Vos, J. C., van Luenen, H. G. A. M., and Plasterk, R. H. A. (1993). Characterization of the *Caenorhabaditis elegans Tc1* transposase *in vivo* and *in vitro*. Genes Dev. 7, 1244-1253.

Vos, J. C. and Plasterk, R. H. A. (1994). *Tc1* transposase of *Caenorhabditis elegans* is an endonuclease with a bipartite DNA binding domain. EMBO J. 13, 6125-6132.

Vos, J. C., Baere, I. D., and Plasterk, R. H. A. (1996). Transposase is the only nematode protein required for *in vitro* transposition of *Tc1*. Genes Dev. 10, 755-761.

Waddell, C. S. and Craig, N. L. (1988). *Tn7* transposition, two transposition pathways directed by five *Tn7*-encoded genes. Genes Dev. 2, 137-149.

Wiehand, T. W., and Reznikoff, W. S. (1994). Interaction of *Tn5* transposase with the transposon termini. J. Mol. Biol. 235, 486-495.

Yannopopulos, G., Stamatid, N., Monastirioti, M., and Louis, C. (1987). *hobo* is responsible for the induction of hybrid dysgenesis by strains of *Drosophila melanogaster* bearing the male recombination factor 23.5 MRF. Cell 49, 487-495.

Yoda, K., Kitagawa, K., Masumoto, H., Muro, Y., and Okazaki, T. (1992). A human centromere protein, CENP-B, has a DNA binding domain containing four potential α helices at the NH_2 terminus, which is separable from dimerizing activity. J. Cell Biol. 119(6): 1413-1427.

Youngren, S. D. A, Boeke, J. D., Sanders, N. J., and Garfinnkel, D. J. (1989). Functional organisation of the retrotransposon Ty from Saccharomyces cerevisiae: Ty protease is required for transposition. Mol. Cell Biol. 8, 1421-1431.

Yuan, J., Finney, M., Tsung, N., and Horvitz, H. R. (1991) *Tc4*, a *Caenorhabditis elegans* transposable element with an unusual fold-back structure. Proc. Natl. Acad. Sci. USA 88, 3334-3338.